Cancer, DNA Repair and Chromatin Structure

James G D Prendergast

University of Edinburgh

Thesis presented for the degree of Doctor of Philosophy

2007

I declare that this thesis has been composed by myself, and except where otherwise stated, is entirely my own work.

James Prendergast
September, 2007

Acknowledgements

I would like to acknowledge Malcolm Dunlop, Harry Campbell, Wendy Bickmore and Colin Semple for their supervision and support over the course of my studentship. I would also like to acknowledge the members of the Dunlop and Semple labs, particularly Susan Farrington and Albert Tenesa, for their contribution to the research presented in this thesis.

My family, and in particular my mother and brother, also of course deserve acknowledging as well as Mr T who bought me my first computer and who I therefore hold a large part responsible for doing bioinformatics.

Key abbreviations

- EST: Expressed Sequence Tag
- SAGE: Serial Analysis of Gene Expression
- CALFC: Clone Average Log_2Fold Change
- BLAST: Basic Local Alignment Search Tool
- LD: Linkage Disequilibrium
- SNP: Single Nucleotide Polymorphism
- ESE: Exonic Splice Enhancer
- ESS: Exonic Splice Silencer
- dS: Rate of divergence at synonymous sites
- dN: Rate of divergence at nonsynonymous sites
- ARCTIC: Assessment of Risk for Colorectal Tumors in Canada
- SOCCS: Study of Colorectal Cancer in Scotland
- COGS: Colorectal cancer Genetic Susceptibility Study

Abstract

Colorectal cancer, the most common cancer in males and females who do not smoke, is diagnosed in approximately 3,500 Scots each year. Despite having a large environmental contribution, the substantial genetic basis of colorectal tumours is still poorly understood. In this project we have adopted a number of approaches to try and further characterise this genetic contribution of colorectal cancer.

To begin to understand tumour progression, we first characterised the gene expression changes observed in various tumours using SAGE, EST and microarray data. Although many genes were identified as differentially expressed in cancers, little congruence was observed between tumour types and even expression platforms.

We next compared gene expression changes observed along chromosomes to local chromatin structure, and showed that regions of constitutively open structure generally show an increase in gene expression in cancer. Despite the lack of congruence between expression data shown previously, we illustrated that such a correlation between gene expression change in tumours and chromatin structure can be observed using various expression platforms and across a variety of tumours.

To further characterise the role of chromatin structure in tumours, we also investigated the rates of mutation and selection across chromatin categories. DNA damage and repair is a key process in cancer progression and we have shown, through inter species alignments, that although chromosomal regions of a relatively more open chromatin structure undergo lower rates of mutation, levels of purifying selection on synonymous sites are highest in regions of closed chromatin.

As part of the COGS/SOCCS group the role of DNA repair in colorectal cancer was finally further investigated through a case-control association study. Tagging SNPs in genes predicted to be associated with DNA repair were selected and subsequently typed by the group in approximately 1000 cases and 1000 controls. The nature of SNPs with evidence of an association with colorectal cancer was finally characterised.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Colorectal Cancer

1.1.1 Background

In 2003 1,858 Scottish males and 1,507 Scottish females were diagnosed with colorectal cancer and approximately 5% of all Scots will develop the disease at some point in their lifetime. Despite dramatic reductions in mortality rates in recent years only around half of these individuals are expected to survive five years [1]. Worldwide, over a million cases of colorectal cancer will be diagnosed this year, accounting for over 9% of all new cancer cases, a figure only exceeded by cancers of the lung and breast. Rates of colorectal cancer incidence are not however constant across the globe with two-thirds of all cases of colorectal cancer occurring in the most developed countries. Survival rates are also markedly lower in the UK than in other Western European countries, and this has been tentatively attributed to delays in treatment and later stages at diagnosis [2].

The diagnosis of tumours at an early stage has however been shown to have a dramatic affect on colorectal cancer survival rates. The Dukes and Modified Dukes systems for staging colorectal tumours as well as the TNM (Tumour, Node Metastasis) staging systems, each defines a tumour according to factors such as degree of invasion of the mucosa, progression to lymph nodes and metastasis. Patients diagnosed at an earlier stage have a markedly improved prognosis than those with a more developed tumour, with 83% of patients diagnosed at Duke's stage A (tumour penetrated into mucosa of bowel wall only) surviving five years compared to only

3% diagnosed at Duke's stage D (tumour spread to multiple organs) [3]. Diagnosing tumours early on in their development can therefore dramatically increase survival rates, and consequently the development of genetic tests for the identification of individuals at risk from the disease is a priority in colorectal cancer research. Likewise understanding the progression of tumours should assist in the development of more affective interventions. There is consequently a need to further understand the genetic mechanisms underlying colorectal cancer.

1.1.2 Molecular progression

Colorectal carcinogenesis has often been described as a multi-stage process involving progressive genetic alterations along a relatively well characterised phenotypic pathway [4, 5]. This step-wise progression from normal to tumour tissue is defined by the adenoma-carcinoma sequence, whose first stage is characterised by the alteration of normal intestinal crypts to aberrant crypt foci (ACF) or microadenomas, often as a result of *APC* loss or mutations in the *k-Ras* gene. The size, number and features of these aberrant crypt foci are in turn associated with the number of benign adenomas that characterise the next stage in tumour progression. Postmortem studies have shown that in Western populations the incidence of these adenomas is approximately 30-40% [4]. Upon the accumulation of certain genetic alterations these adenomas are subsequently transformed into in situ tumours, and in the final stage to metastatic carcinomas [6]. It should be noted however that this simplified pathway of tumour progression is not applicable to all cases of colorectal cancers. For example carcinomas have been shown to arise from lesions rather than adenomas [7] and follow the inflammation-dysplasia-carcinoma sequence. However the types and frequencies of mutations observed in these alternative routes of carcinogenesis are believed to be analogous [8].

One of the earliest molecular events in the progression of almost all colorectal tumours is loss of function of the *APC* gene product. Mutations in *APC* have been identified in up to 80% of sporadic colorectal tumours and germline mutations in the gene lead to familial adenomatous polyposis [4]. The molecular events that characterise the later stages of colorectal carcinogenesis can broadly be broken down into two alternative pathways; microsatellite instability (MSI) and chromosome instability (CIN) (although tumours with near-diploid chromosomes displaying no microsatellite instability can be observed, they are relatively rare [9]).

The microsatellite instable phenotype results from defects in DNA repair. Maintaining the integrity and accuracy of DNA and the genes it encodes is an important step in avoiding cancer progression and a number of the genes associated with colorectal cancer have been shown to be involved in DNA replication and repair. These include various mismatch repair genes. These genes, involved in the correction of single-base mismatches that arise during DNA replication, also eliminate insertion deletion loops that result from gains or losses of short repeat units within microsatellite sequences. Defects in these genes lead to the accumulation of detectable abnormalities at these microsatellite regions, and MSI is consequently a hallmark of mismatch repair (MMR) deficiency. People with a deleterious germline mutation or hyper-methylated promoters at a MMR gene display large numbers of instable loci (microsatellite instability-high; MSI-H) and this is frequently seen in patients with heriditary non-polyposis colorectal cancer (HNPCC). Around 10 to 15% of sporadic colon tumours will also show at least some microsatellite instability (MSI-L) [10, 11, 12]. Although the development of tumours after the loss of efficient mismatch repair is often simply the result of an accumulation of mutations in key genes such as *TP53*, some tumour suppressors, such as *TGFBR2* and *BAX*, contain microsatellites within their coding regions. Loss or gain of nucleotides at these microsatellites, resulting from this loss of efficient MMR, will lead to frameshifts inactivating the corresponding gene [13, 14].

The CIN phenotype is characterised by the gain or loss of chromosomal regions or even entire chromosomes and is observed in approximately 85% of colorectal tumours. Although the mechanism of CIN is poorly understood it has been associated with various mitotic-spindle and DNA replication checkpoint genes. For example mutations in two genes that control the human mitotic checkpoint, *BUB1* and *BUBR1*, have been observed in several CIN colon cancer cell lines. Although it is suggested that the CIN and MSI phenotypes may be a result of cancer rather than its cause, examination of colorectal and endometrial cancers illustrates that there is an inverse relationship between CIN and microsatellite instability, i.e. MMR deficient cancers generally do not display chromosome instability and vice versa. The rate of tumour progression is also faster in MSI cancers and it is therefore likely that CIN and MSI are alternative routes to tumour development [14, 15].

1.1.3 Major syndromes

It has been shown in numerous studies that individuals with a first-degree relative diagnosed with colorectal cancer are two to three times more likely to develop colorectal cancer than the general population [16, 17, 18]. This relative risk is increased dramatically if two or more relatives have been diagnosed with colorectal cancer, and especially if the onset of the disease was relatively early in their lifetimes. Where a number of individuals from the same family are affected with the disease in this way, there is strong evidence that a genetic syndrome may be responsible. Of the major colorectal cancer syndromes Familial Adenomatous Polyposis (FAP) displays the highest absolute lifetime risk, with 90% of affected individuals developing colorectal tumours by the age of 45 [19]. The gene responsible for the majority of FAP cases is *APC*, and mutations in this gene lead to individuals developing hundreds to thousands of adenomatous polyps in their colon and rectum that can each subsequently develop into tumours. As FAP results from dominant mutations in the *APC* gene and offspring of affected individuals are at a 50% risk of developing the disease, FAP is a strong candidate for genetic screening within affected families [20].

A number of FAP patients (i.e. patients with multiple colorectal adenomas) do however have no identifiable pathogenic germline mutations located in the *APC* gene ORF (open reading frame) [21], and it has been shown that in approximately 7-8% of affected individuals the FAP phenotype is caused by biallelic mutations at the *MYH* gene (*MYH*-associated polyposis) [21, 22, 23]. MYH is involved in repairing 8oxoG:A mispairs that primarily result from oxidative damage. As oxidative damage is particularly prevalent within the gut, inactivation of *MYH* is likely to make carriers particularly susceptible to mutations that lead to colorectal tumours. Biallelic *MYH* mutations are however relatively rare within the general population [24].

Hereditary Nonpolyposis Colorectal Cancer (HNPCC or Lynch Syndrome), like *APC* associated FAP, is an autosomal dominant disorder with high penetrance. 80% of affected individuals develop tumours by the age of 75 [25]. However, unlike FAP patients, individuals with HNPCC do not display an abnormal number of colorectal polyps. Consequently cases of HNPCC are more difficult to identify, especially as the disorder can also lead to a number of other cancers complicating an individuals family history. HNPCC results primarily from defects in the mismatch repair pathway and significantly more affected individuals with this syndrome display evidence of microsatellite instability than is observed in all colorectal cancer patients

[26]. Although there are at least 18 human genes associated with DNA mismatch repair, defects in the *MLH1* and *MSH2* genes account for the majority of HNPCC cases. Redundancy among the other 16 genes may account for their relatively low contribution to the disorder [27, 11, 10].

Although together FAP and HNPCC account for the vast majority of cases of inherited colorectal cancer in strongly affected families, there are a number of further rare, inherited syndromes associated with colorectal tumours that include Peutz-Jeghers syndrome (PJS) and juvenile polyposis syndrome (JPS). Peutz-Jeghers syndrome is an autosomal dominant disorder that is characterised in affected individuals by multiple hamartomatous and adenomatous gastrointestinal polyps as well as melanin spots on the lips and buccal regions [28]. It is estimated that 93% of individuals with PJS will develop a noncutaneous cancer by the age of 64 with 39% of them developing a tumour of the colon [29]. The majority of PJS cases have been shown to be associated with germline mutations at the gene encoding the serine-threonine kinase, *STK11* [30]. Juvenile polyposis syndrome, that is also an autosomal dominant disorder, is a rare childhood-onset disease demonstrating with hamartomatous polyposis, diarrhea and gastrointestinal tract bleeding. This disorder has been linked with the *SMAD4* and *BMPR1A* genes in up to 60% of cases [31, 32].

1.1.4 Major tumour suppressors and oncogenes

In 1986 Herrera et al. identified a deletion of 5q in a patient with familial adenomatous polyposis [33]. This initial discovery eventually led to the identification of one of the key genes in colorectal cancer, the tumour suppressor *APC*. As mentioned, mutations in *APC* have been identified in up to 80% of sporadic colorectal tumours and certain germline mutations in the gene lead to familial adenomatous polyposis. Further mutations in *APC*, such as I1307K, have also been potentially associated with familial colon cancer. These mutations, unlike those underlying FAP, are not thought to directly underlie the development of colorectal tumours despite families carrying these mutations displaying elevated levels of colorectal cancer incidence. Rather it is believed that these mutations act by promoting sporadic mutations at other neighbouring parts of the gene. The T to A transversion underlying I1307K leads to an unstable run of eight adenines (from $(A)_3T(A)_4$) and a somatic mutation at this poly-adenine tract is observed in approximately 42% of tumours involving I1307K carriers, compared to only 4% of controls (127 cases and 127 controls) [34].

The majority of these mutations are an insertion of a single adenine leading to a frameshift and mis-translation of downstream codons. *APC* is consequently a key gene in colorectal cancer and as a result has come to be known as the "gatekeeper" gene [35, 36].

The mode of action of APC is through the modulation of the β-catenin protein. In its wild-type form the APC protein inhibits β-catenin. However certain mutations in the *APC* gene can lead to its uncontrolled accumulation in the cytoplasm ultimately leading to aberrant cell signal transduction and growth. It is believed this occurs through the association of β-catenin with certain T-cell transcription factors, leading to aberrant expression of key Wnt responsive genes (e.g. *c-myc*, *c-jun*, *Fra* and *cyclin D1*) [35, 37].

In 1988 Volgelstein et al. showed that a second region, located on chromosome 18, was lost in approximately 73% of colorectal carcinomas [38]. The *DCC* (deleted in colorectal carcinoma) gene was later identified in this region and it was shown that the expected survival time of patients was significantly reduced if this gene was found not to be expressed. However recent evidence suggests that the *DPC4* (*SMAD4*) gene, found 1.26Mb upstream of *DCC*, may also be a key gene from this region [36]. Of 20 unrelated individuals with juvenile polyposis syndrome sequenced at this gene, 5 were shown to carry the same exonic 4bp germline deletion. None of 240 controls examined carried the same mutation [31, 39].

Another tumour suppressor involved in colorectal cancer progression is *p53*. Up to 75% of sporadic colorectal tumours contain a mutation in the *p53* gene, and affected patients have a significantly poorer prognosis in both terms of outcome and survival time. Wild-type p53 is involved in DNA repair via the arrest of the cell cycle at G1 (allowing corrections of the DNA to be made) as well as the induction of apoptosis. Consequently mutations of the *p53* gene can lead to the accumulation of DNA damage as well as uncontrolled cell growth [36].

One of the apparently key oncogenes involved in colorectal carcinogenesis on the other hand is *KRAS*. Activation of *KRAS* leads to the continuous transmission of extracellular growth signals to the nucleus resulting in the uncontrolled growth of affected cells. However the Src non-receptor tyrosine kinase, *c-myc* and *c-erbB2* have all also been associated with the colorectal cancer phenotype [36, 40].

1.1.5 Low penetrance genes and polymorphisms

The major genetic syndromes discussed in 1.1.3 are thought to only account for 2-6% of colorectal cancer cases [41], despite twin studies suggesting that up to 35% of cases have some form of heritable contribution [42]. In agreement with this, a substantial number of colorectal cancer cases have been shown to be familial in nature, i.e. cases are observed between related individuals more often than would be expected by chance, but in such a way that they can not be explained by the patterns of inheritance associated with the FAP or HNPCC syndromes [36]. Relatives of patients with supposedly sporadic cases of colorectal cancer have also been shown to be at an increased risk of developing the disease than the general population. There is therefore a substantial genetic contribution to colorectal cancer not accounted for by the high penetrance genes associated with the major genetic syndromes, and although there may still be further high penetrance genes to be discovered, it is likely that combinations of lower penetrance alleles account for most of this unknown genetic contribution [41].

Despite substantial effort in recent years to identify low-penetrance genes involved in colorectal cancer, few genes have been unequivocally associated with the disease. The low penetrance nature of the genes means they will rarely cause multi-case families such as those observed with FAP or HNPCC and consequently case control studies, where the frequencies of polymorphisms in affected individuals are compared to those in unaffected controls, have become the method of choice for detecting these alleles. However, few polymorphisms shown to be associated with colorectal cancer have been replicated across studies and this is at least partly the result of the insufficient sample sizes and the lack of sufficiently rigorous statistical testing used in these analyses. Variants that have however been shown to be significant in more than one study include a rare VNTR (Variable Number of Tandem Repeats) allele in the *HRAS-1* oncogene, the polymorphism associated with the rapid acetylator phenotype of the N-acetyltransferase *NAT2* and a synonymous change in the *MTHFR* gene involved in folate and methionine metabolism [43, 41, 44]. Models of the contribution of low penetrance alleles to the onset of colorectal tumours have however predicted there may be hundreds of rare polymorphisms with some form of association with the onset of colorectal cancer.

1.1.6 Genetic testing/screening

The use of screening techniques such as faecal occult blood testing (FOBT) and sigmoidoscopy have been shown to reduce colorectal cancer mortality rates by between 15 and 50%, and the NHS has recently announced that from 2006 all individuals aged between 60 and 69 will be given the opportunity to be tested for colorectal cancer using FOBT. This technique, that detects small amounts of blood in a patients faeces that may result from polyps or tumours, although less effective than sigmoidoscopy or colonoscopy is less invasive and has been shown to display the highest participation rates [45]. Individuals from high risk families such as those with HNPCC or FAP are also offered regular screening (generally via colonoscopy). However faecal blood detected by FOBT is not necessarily an indication of a tumour; likewise FOBT and sigmoidoscopy have been shown to miss a substantial number of tumours and a negative test is not a guarantee that a patient is free from colorectal cancer. It has therefore been proposed that genetic testing should precede or be used in conjunction with these screening procedures. For example, 50% of the offspring of a single FAP patient will not carry the *APC* mutations associated with the disorder and therefore may prefer to not undergo regular invasive screening if genetic tests prove this is the case. Likewise the genetic screening of the general population for mutations associated with colorectal cancer could be used to identify high risk individuals for whom regular screening would be particularly beneficial. For this type of genetic testing to be cost effective however, mutations that confer a substantial relative risk and that are relatively common in the population need to be identified.

1.1.7 Environmental factors

As mentioned, the incidence of colorectal cancer is higher in developed countries than in the developing world and this has been attributed to the differences in diet and other environmental factors observed between these regions. Various studies have shown that the incidence of colorectal cancer in immigrants to the USA from countries such as Japan, whose incidence of colorectal cancer is relatively low, can increase rapidly to match or even surpass that of the host population [46]. There is therefore strong evidence that dietary and environmental factors contribute substantially to the development of colorectal tumours. However the evidence as to what factors affect the rates of colorectal cancer is often weak and contradictory, a result of the complex interplay of many dietary, genetic and environmental factors and the

difficulty of isolating the specific contributing factor.

One of the factors with the strongest links to colorectal tumours however are Nonsteroidal Antinflammatory Drugs or NSAIDs. NSAIDs, such as sulindac and aspirin, have consistently been shown to have a protective effect with respect to cancer development. A study by Chan et al. in 2005 showed that those women who consumed more than 14 325-mg tablets of aspirin a week for at least ten years had a 53% decreased risk of colorectal cancer compared to women who took no aspirin [47]. Similar affects have been observed with other NSAIDs [48]. Although these results initially appear promising, large doses of NSAIDs have been associated with side-affects such as an increased risk of gastrointestinal bleeding, and consequently the potential benefits from the long term use of high doses of NSAIDs are outweighed by their inherent risk [48] (it has been predicted that the prevention of 1 or 2 colorectal cancer cases per 10000 people will be at the cost of 8 cases of severe gastrointestinal bleeding). Further understanding of the mechanisms by which NSAIDs protect against colorectal tumours may however lead to better treatments in the future.

Another factor potentially associated with colorectal cancer prevention is Hormone Replacement Therapy or HRT. A meta-analysis by Nanda et al. of a number of observational studies concluded that recent use of HRT conferred a 33% reduction in colon cancer risk (the protective affect of HRT appears to be lost after the cessation of HRT treatment) [49]. This link between HRT and colorectal cancer is further (tentatively) supported by the fact that the incidence of colorectal cancer among females has fallen since the introduction of HRT but remained relatively stable among males. A causal link between HRT and colorectal tumour prevention has however yet to be established and certain randomised controlled trials have found no supporting evidence of a link between HRT and colon tumours. Given the potential side-affects of sustained HRT, such as breast and endometrial tumours, the use of estrogen itself is unlikely to be a viable treatment of colon cancer.

HRT has also been predicted to modify the association between body weight and colorectal cancer risk. An association between body weight or BMI and colorectal cancer has been observed in men in a number of studies but no such association has been observed in women, and HRT has been proposed as an explanation of this discrepancy [50]. Direct links between the intake of dietary factors such as fat, fibre, vitamins or antioxidants are however generally weak or contradictory and further work is required to elucidate the impact of diet on colorectal cancer incidence.

Other environmental factors with weak or contradictory evidence of modifying

colorectal cancer risk include smoking, alcohol and physical activity [51]. Consequently, despite the environment playing a large part in the progression of colorectal tumours, the interplay between factors is likely to be complex and has yet to be fully elucidated.

1.2 Bioinformatics and Cancer

Primarily due to the increase in large scale digital datasets, the roles of bioinformatics and computational biology in oncology have expanded dramatically in recent years. The advent of the genomic era has moved the focus of cancer research increasingly to the global level and to the integration and analysis of complex datasets. This is perhaps most apparent in the field of gene expression where there has been an explosion in genome-wide analyses of tumour transcript levels. The role of bioinformatics has however extended to all areas of cancer research and some examples of this are discussed below.

An area in which bioinformatics has played a major role is cancer diagnostics. The aim of cancer diagnostics is to identify and stage tumours, primarily through the detection of specific proteins, RNA or other molecular markers in an affected individual. For example, cancerous prostate cells produce abnormally large amounts of the prostate-specific antigen (PSA) into the bloodstream, consequently a number of tests have been developed that measure the levels of this protein in patients [52]. The initial identification of these molecular markers requires the comparison of tissue/blood samples from normal and affected individuals and to identify protein diagnostic signatures in the bloodstream mass spectrometry is commonly used. However, the high dimensionality, substantial noise and complex spectra that characterise mass spectrometry results means that mass spectrometry data requires substantial computational processing to piece together the most parsimonious protein contents of various tissue samples [53]. For example the sequencing of the human genome has allowed the mass spectra of all potential proteins to be determined [54]. Likewise, the characterisation of protein profiles that characterise tumour samples also generally involves computational analyses such as those based on support vector machines (SVM)[55].

Cancer diagnostics is not however restricted to the analysis of protein levels, transcript levels have also been used to characterise tumours. For example bioin-

formatics analysis of genome-wide gene expression studies has shown that certain tumours can be sub-classified according to their expression profiles [56]. The role of bioinformatics in gene expression and cancer is discussed further in chapter 2.

Bioinformatics has also been used extensively in disease gene identification. The identification of genes associated with a disease traditionally involves the use of linkage or association studies. Not only are bioinformatic tools generally used in the design of these studies (for example in picking tagging variants [57]) but as both these techniques are limited in their resolution, so that often a number of potential candidate genes remain, a number of bioinformatic approaches have been developed to prioritise genes in candidate regions. The most common first approach to ranking candidates is to identify if any of the genes of interest have a known function associated with the disease. This is most commonly achieved through Gene Ontology (GO) term annotation. The gene ontology project is an attempt to assign certain terms describing biological processes, cellular components and molecular functions to all genes and gene products. A gene in a candidate region that is involved in a key process associated with the relevant disease, is generally thought more likely to be the gene of interest from that region. Programs that have been developed around this principle include Freudenberg and Propping's disease clustering algorithm [58], G2D [59], and POCUS [60]. (Gene Ontology terms, tag selection, association studies and cancer are discussed further in chapter 5)

Gene Ontology terms and other annotations are also often used in conjunction with expression data to identify disease genes (e.g. SUSPECTS [61]). For example, although a number of genes may be differentially expressed in a particular disease, only a subset of these will also be associated with key biological processes or pathways. Further annotations that have been used in prioritising candidate genes in this way include protein interaction data, domain annotations, PubMed reports and eVOC annotations [62].

It has also been shown that disease genes may, as a group, have certain sequence characteristics that mark them out from the rest of the genome. For example, Lopez-Bigas and Ouzounis illustrated that proteins involved in hereditary disorders tended to be longer, more conserved, phylogenetically extended and without close paralogues [63]. Likewise CpG islands, long 3' UTRs, and large numbers of exons have been associated with hereditary disease genes. Consequently programs such as PROSPECTR [61] have been developed that allow the prioritisation of potential disease genes according to their sequence characteristics.

The recent rapid expansion in sequenced genomes has also further assisted in the examination of candidate disease regions. Not only can disease genes/regions and other annotations identified in model organisms be more readily transferred to the human genome, but multi-species conservation can be a key tool in further refining regions associated with the disease. For example, a number of ultra-conserved, non-genic regions of the human genome have been potentially associated with disease [64, 65]. Likewise conservation can be used to identify domains and regulatory motifs within a region that may be involved in disease progression. Conservation has also been used, primarily in monogenic diseases, to identify variants potentially associated with disease. For example those non-synonymous changes in a gene that are at a position that is particularly conserved across species are thought more likely to be deleterious, particularly if the observed change is not observed in other homologous transcripts [66]. Further criteria that have been used to predict potentially deleterious polymorphisms include the affect of a polymorphism on splicing, domain score predictions and protein structure [67]. SNP (single nucleotide polymorphism) prioritisation is discussed more in chapter five.

The ever increasing number of sequenced genomes has also meant that bioinformatics has begun to play an ever more important role in drug target identification and validation. Until recently drugs were tested individually and generally influenced a previously known target. However the advent of the genomic era has led to the acceleration of the drug discovery process. The first step in drug discovery, the identification of a potential drug target, often involves identifying family members and homologues of previously known drug targets and the subsequent classification of these genes and their products according to their structural properties, locations in the cell, evolutionary history [68] and expression profiles in both the target and other organisms [69]. Once a group of potential targets has been identified, and the molecular basis of the disease elucidated, it is then possible to predict computationally, drugs that will interact in an appropriate fashion with the target protein. If an appropriate structure of the drug target has yet to be determined, a number of algorithms have been developed that attempt to predict a protein's structure from its amino acid sequence and its homology to proteins of known structure. Once a suitable drug compound is determined, its activity is often further refined through the computational analysis of the precise molecular structures responsible for the drugs activity and potency.

The role of bioinformatics in cancer epigenetics is also expanding rapidly. Un-

til recently there has been little data available on genome-wide epigenetic states in the human cancer genome. However the dramatic reduction in costs of sequencing and array technologies have led to a number of studies investigating the methylation patterns across the human genome. Although some studies have involved the computational analysis of large bisulphite sequencing datasets [70, 71], genotyping microarrays have allowed the investigation of epigenetic silencing on a truly global scale. The underlying principle behind these studies being that where a copy of a gene is silenced, all alleles corresponding to the transcripts from that gene will no longer be expressed. In a similar fashion genotyping platforms have been used to determine copy number variation in the human genome through looking at the ratio of allele intensities obtained from DNA alone [72].

These however are only some of the many areas in which bioinformatics is playing a key role in cancer research and given the wealth of data being generated in this field, the role of bioinformatics is only likely to increase in the future.

Chapter 2

Tumour Gene Expression

2.1 Introduction

The link between cancer and gene expression has been well characterised. Several of the most important genes in cancer progression can be classified as oncogenes or tumour suppressors, whose aberrant expression through inappropriate methylation, RNA degradation, mutations or copy number changes, can lead to tumour development. Although further analysis is generally required, as it is necessary to determine those genes whose change in expression is the cause of the relevant tumour and not simply its result, expression analysis can be a valuable tool for highlighting potential candidate cancer genes.

The characterisation of gene expression has also proven a useful tool in identifying those genes whose expression level is specific to a certain tissue or disease state, analyses that can potentially allow for the development of specific therapies and diagnostic tests. For example Alizadeh et al. [56] used gene expression profiles to define two molecularly distinct sub-types of diffuse large B-cell lymphoma, likewise van't Veer et al. [73] identified gene expression signature in breast cancer biopsies associated with patients with a relatively poor prognosis.

Many genes associated with cancer also belong to one of several key pathways, for example apoptosis, the cell cycle or DNA repair, likewise many of the key cancer genes have been shown to interact [74]. Analysis of expression signatures across tissues can be used to identify those genes whose expression profiles are similar to genes previously associated with the disease [75] and which themselves could be involved in cancer progression.

Further uses of gene expression analysis have included the identification of non-specific expression of a drug target that may result in undesirable toxicity and the characterisation of copy number changes in tumours [76]. The investigation of tumour gene expression, and its comparison to expression levels in normal tissue, is consequently a vital tool in further understanding the development and progression of colorectal cancers.

In this analysis we investigated the expression of genes in normal and tumour tissue samples via the analysis of Expressed Sequence Tags (ESTs). Our aim in this study was to not only identify genes differentially expressed in various tumour types, but to also identify differently expressed genomic regions, genes expressed primarily in particular tissues of interest and gene coexpression. These data could then subsequently be used in the prioritisation of genes for further examination in association studies of colorectal and other cancers. For example, which genes are expressed specifically in the colon and also show differential expression in colorectal tumours and may therefore be specifically associated with colorectal cancer? What genes are coexpressed with known tumour associated genes? Also what genes are differentially expressed in not only colonic tumours but also other forms of cancer?

Out of interest we were also keen to compare these EST results to those from other expression platforms. If we are to prioritise genes for candidate association studies using expression data what difference does it make if other expression platforms are used? Does the list of candidates dramatically change? What is the overlap between platforms? Perhaps those genes that are consistently differentially expressed across platforms are the true positives and therefore the truly interesting candidates.

2.1.1 Expressed Sequence Tags

ESTs are short, single pass reads of randomly selected cDNA clones that are a valuable resource for the investigation of tumour gene expression. Although initially used to identify the genes present in a genome of interest [77] they have been increasingly used to characterise transcript expression profiles. Megy et al. for example identified a number of genes overexpressed in the heart through the comparison of cardiac and non-cardiac derived ESTs [78]. Of the 35 genes deemed differentially expressed, five had previously been associated with cardiac disorders and one was located in the locus of a bleeding disorder. Expressed sequence tags can therefore play an important role in elucidating potential disease gene candidates.

ESTs are produced by first converting the mRNA content of a sample to cDNA via reverse transcription. These more stable cDNA libraries are then typically cloned into plasmid vectors and then sequenced from the 5' or 3' end to produce ESTs that are generally a few hundred base pairs in length. It is important to note that although 3' ESTs by and large match the 3' of a transcript due to the reverse transcription being primed by an oligo-dT sequence from their 3' end, the resulting cDNAs are often incomplete and can lack a corresponding 5' end [77]. ESTs sequenced from the 5' end of a cDNA may therefore not necessarily represent the 5' end of the mRNA but rather may derive from a more central portion of the gene.

The proportion of ESTs in a library matching a particular transcript should be related to the levels of that transcript in the corresponding sample; as those transcripts whose levels are high in a sample should be cloned into more vectors and consequently sequenced more often. The use of ESTs to determine transcript levels in this way is however dependent on the ability to unambiguously match ESTs to the transcripts from which they were derived. The sequencing of ESTs can often be of a poor quality and the large number of paralogues and homologous regions within the human genome can make assigning ESTs to individual genes difficult. Likewise many EST libraries have undergone normalisation or subtraction to amplify the representation of genes of low expression. These libraries are therefore unsuitable for use in gene expression studies as the link between EST numbers and gene expression has been disrupted.

2.1.1.1 Statistical analysis

As most EST studies are unreplicated determining the significance of any observed change in EST numbers between samples requires the use of appropriate statistical tests. Although Fisher's exact and chi-square tests are often used in conjunction with a Bonferroni multiple test correction in EST studies [79], this type of analysis is likely to be too conservative and exclude many genes that display true biological changes in expression. Likewise, the use of the Fisher's exact test when both the row marginal totals and column marginal totals are not fixed is controversial (in EST studies only the number of clones sampled in each library is fixed) and therefore Audic and Claverie developed a new statistical test specifically for use in tag based gene expression studies [80]. The sole assumption of this test is that the observation of any particular cDNA is rare, which is the case in most libraries where each cDNA

is represented by at most only a few percent of all ESTs.

Whatever statistical test is used in a gene expression study it is necessary to account for the large number of separate tests performed and to try and minimise the number of false positives. Traditionally multiple testing has often been corrected for by applying a Bonferroni correction, so that the probability of observing a single false positive among all genes is less than 5%, however this is likely to be far too conservative for gene expression studies, especially as the cost of a false positive in a gene expression study is relatively low (individual gene expression levels can be easily confirmed). An alternative to the Bonferroni correction is to calculate the q value for each gene [81]; a measure based on the false discovery rate (rather than the false positive rate as in traditional p values) that takes into account the number of features being tested. The q value of a test represents an estimation of the number of false positives that will be obtained if the respective test is called significant. Not only can q values be thought of as less conservative than Bonferroni corrected p values but they are also more informative as they predict the number of genes among a given significant set that are likely to in fact show no difference in their gene expression. A q value significance cutoff can consequently be set according to the number of these genes that one is willing to accept.

2.1.2 Other expression platforms

In this analysis we attempted to confirm those genes deemed differentially expressed in our study of EST libraries through comparisons with datasets derived on other genome-wide expression platforms. Although each individual expression platform has limitations and will generate a set of false-positives, we hypothesised that those genes shown to be differentially expressed across platforms and samples are more likely to be truly differentially expressed. We therefore compared the results from our EST analysis to those derived from SAGE (serial analysis of gene expression) and microarray studies.

2.1.2.1 Serial Analysis of Gene Expression

The SAGE technique for determining gene expression was first developed within the Vogelstein-Kinzler colon cancer laboratory at the Johns Hopkins Oncology Centre [82]. Like techniques based on ESTs, the SAGE method involves isolating unique sequence tags from transcripts present within a tissue/sample, whose relative abun-

dance should approximate to the expression levels of the corresponding mRNA. This is achieved through a number of steps. First, the poly-A tails of mRNAs within a sample of interest are captured on oligo-dT beads, and these captured mRNAs are reverse transcribed to cDNA with the oligo-dT acting as a primer. To isolate unique tags from each mRNA species the anchored cDNAs are then digested with a restriction enzyme, usually NlaIII, so that only the bases following the final cutting site of this enzyme are left attached to the magnetic beads. By attaching linkers that contain the recognition site of a type II restriction enzyme, such as BsmFI, to the resulting sticky ends, the ten base pairs immediately following the NlaIII site can be isolated. To determine the mRNA content of the sample of interest these resulting tags are then finally sequenced. This is achieved by first ligating pairs of tags tail to tail to form ditags, amplifying these ditags by PCR (polymerase chain reaction) and then releasing the linkers using NlaIII. The concatenation, cloning and sequencing of these ditags then allows the expression levels of each mRNA species to be estimated.

This SAGE technique for determining gene expression generally has the same advantages as methods based on ESTs; namely the results are digital in nature, no pre-selection of genes is required and results are highly comparable between experiments and labs. The number of short SAGE tags per library are however generally substantially higher than the number of (longer) ESTs produced per experiment, and consequently SAGE can be more powerful at detecting genes with lower expression levels. There are however a number of drawbacks to the SAGE technique, primarily because this method for determining gene expression is strongly dependent on being able to unambiguously assign a unique tag to each mRNA, which in practice is often not possible. For example, some mRNAs simply do not contain the appropriate restriction site, or the final site within the mRNA is immediately upstream of the poly-A tail. Likewise imperfect digestion will lead to some tags being derived from sites further upstream than the final site so that multiple tags correspond to the same gene. Sequencing error and polymorphisms in these short 10bp tags will also lead to some being assigned to the wrong transcript, and tags are often generated that are not 10bp in length making the boundaries within ditags difficult to determine. Blunt end ligation used within the SAGE protocol is also dependent on the corresponding terminal bases so that tag generation can be biased. Although protocols have been determined to try and minimise these problems; for example repeating the experiment with different restriction enzymes, producing longer tags and excluding tags with low counts, these changes largely diminish the advantages of the SAGE

technique [83].

The inherent problems of the SAGE technique can be illustrated using data from one of the largest SAGE gene expression studies carried out to date (and the one used in our analysis) that characterised the gene expression profiles within normal human colorectal epithelium, colorectal tumours and pancreatic cancers [84]. Of over 300,000 tags generated by this study over 70,000 were unique and 86% of transcripts were represented by less than five tags. Consequently few genes can be deemed significantly differentially expressed from this data, as there are simply insufficient tags to provide the necessary power.

2.1.2.2 Microarrays

Microarrays have become increasingly popular for determining genome-wide expression profiles, primarily because of their ease of use. Determining the expression profiles in a variety of tissue samples can be done relatively quickly and at a relatively low cost. Determining gene expression through the use of SAGE or ESTs on the other hand can require a substantial level of optimisation and sequencing.

The technology behind microarrays evolved from that of the Southern blot, where the presence or absence of specific DNA sequences on a membrane is determined through the addition of complementary, labeled probes. If the sequence of interest is present, it will be bound by the probe and can consequently be visualised via its corresponding label. The approach of microarrays is not however to add individual probes to a DNA sample but rather to add the DNA or RNA sample to a collection of thousands of probes, each of which corresponds to the complement of a section of a transcript that may be expressed in the tissue of interest. By labeling the RNA sample that is hybridised to the microarray with fluorescent dyes it is possible to determine the expression levels of individual transcripts; as the more transcripts that bind to the corresponding probes the greater the total fluorescence. In practice however the relationship between transcript levels and fluorescence is not linear, in part due to intensity-dependent effects, and microarray data must undergo substantial processing.

The main stages of microarray processing are scanning/image analysis, normalisation and transformation. The image analysis stage of microarrays involves determining the fluorescence level of each spot/probe set on each array. In Affymetrix and other types of arrays these intensities are then corrected for non specific hybridisa-

tion by adjusting them according to the intensity of corresponding mismatch probes. As probe intensity levels may differ between arrays for a number of non-biological reasons (differences in RNA quantity, differences in labeling efficiency, scanner setup etc) to allow for the comparison of intensities between arrays (or within two colour arrays) it is necessary to normalise these intensities. A large number of different normalisation procedures exist including loess, rma (robust multi-array) and vsn (variance stabilising normalisation) but perhaps the simplest example is global median normalisation in which the intensities on each array are simply scaled so that the median intensity on each array is the same. Different normalisation procedures are available as each makes different assumptions about the data and required outcome, for example the choice of normalisation procedure can depend on whether or not the expected levels of certain probes are known or whether the distribution of expression levels is expected to be constant. Before or after normalisation microarray data are often transformed to improve comparability and/or signal to noise ratios and the simplest example of this is the log transformation. Microarray error distribution and the advantages and disadvantages of the log transformation are discussed in more detail in chapter 3.

The main problems associated with microarrays are primarily related to this relationship between probe-target binding and fluorescence. For example probes can differ in their hybridisation efficiencies, so that some transcripts are less easily bound by their corresponding probe and can be more easily released during the washing of unbound RNA from the microarray. Likewise the incomplete dissociation of target structures may inhibit target-probe binding. Certain RNAs may also bind inappropriate probes or if a transcript's level is particularly high a probe set can become saturated. Unlike the SAGE and EST techniques comparisons of microarray datasets between experiments and labs is also difficult and it is necessary to pre-specify the content of a microarray so that the expression levels of unknown/unrepresented genes will not be detected. With the advent of large scale sequencing technologies, it is likely that the use of microarrays in global gene expression studies may begin to diminish.

2.2 Methods

329 adult, human EST libraries were obtained from dbEST that were non-normalised, non-subtracted and not obtained using the ORESTES (open reading frame ESTs) technique [85] (although the ORESTES libraries provide a higher coverage of the central, coding regions of transcripts, the approach leads to a certain level of pseudo-normalisation [86]). Selected libraries contained a minimum of 250 sequences and were grouped according to tissue and disease state. All 1,802,156 ESTs were subsequently RepeatMasked (`http://www.repeatmasker.org/`) and an initial attempt made to identify them through Blasting [87] them against the Ensembl cDNA database. Ensembl genes were used in this analysis due to the support at Ensembl for downstream analyses (Ensembl APIs, EnsMart etc). Any EST that matched an Ensembl cDNA sequence with 95% identity over at least 100 base pairs was assigned to that cDNA, a stringent enough threshold to exclude most sequence similar paralogues (any paralogues that are above this threshold are much more likely to lead to the loss of signal than to lead to spurious false positives). This high threshold was used to ensure ESTs were not assigned to closely related paralogues. The remaining 527,951 ESTs were grouped together according to those that shared limited homology with one another (again using the BLAST program as in Megy et al.; score > 40 and e-value $< 10^{-5}$) and formed into contigs using the CAP3 program [88]. A subsequent attempt was made to assign these contigs to an Ensembl gene. Remaining ESTs were discarded [78]. In the analyses in this chapter I focused on those ESTs we were able to assign to a known gene. As in Megy et al. [78] BLAST rather than BLAT [89] was used to assign ESTs to genes, however this is simply a result of the fact that BLAT had only just been released when this analysis was undertaken. BLAT would have undoubtedly been a better choice for this analysis (not least due to its increased speed).

To identify potential differential gene expression between groups of ESTs (e.g. prostate normal and prostate tumour) we utilised the Audic Claverie [80] and chi-squared tests to compare the number of ESTs that had been assigned to a particular gene or contig. The p values derived by these tests were subsequently used to calculate more informative q values through the use of the R statistics package [81]. Length corrections were not required as only ESTs primed from the ends of transcripts were used.

SAGE tag counts for pertinent libraries were obtained from the CGAP website

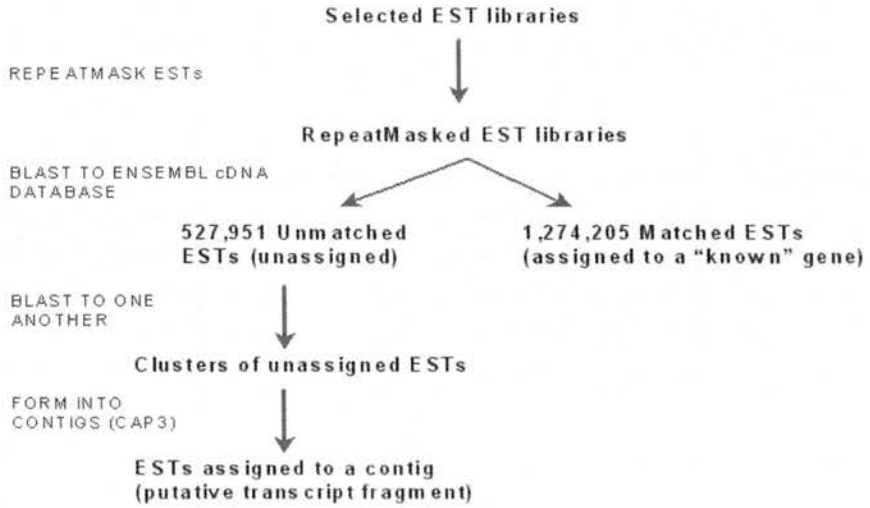


Figure 2.1: Procedure for assigning ESTs to genes or transcripts

[90]. NCBI SAGEmap data [91] was used to map tags to Ensembl genes via LocusLink [92]. Those tags that mapped to more than one Ensembl gene were ignored. As with the EST data, libraries were grouped together according to tissue and disease state and counts were analysed using the same statistical tests.

Microarray data were obtained from the Gene Expression Atlas website for three normal and five cancerous prostate experiments [93]. The average difference values for each probeset on each chip had been scaled by GNF using the standard Affymetrix scaling algorithm (i.e. the top and bottom 2% of probesets ranked by their average intensity after perfectmatch-mismatch background correction were removed and the remainder scaled so that their mean equaled 200). Values less than twenty were set to twenty and averages made for each probe in each tissue type. Probes were mapped to Ensembl gene IDs using data available at the Affymetrix website. Probes that mapped to more than one Ensembl gene were excluded.

2.3 Results and Discussion

Having assigned as many ESTs as possible to either a gene or a contig it was possible to analyse the distribution of their corresponding counts among the different tissues and disease states. Analysis of the results in the corresponding appendix table[1] high-

[1]http://www.hgu.mrc.ac.uk/users/james.prendergast/appendixtable.xls

lights that a number of the genes identified as differentially expressed in colorectal cancer had previously been shown to be differentially expressed in colonic tumours. For example Yow et al. showed that laminin receptor 1 was expressed approximately 9-fold higher in colon tumour tissue than in adjacent normal colonic epithelium [94], a value that is in agreement with our observed change of x8.4. Likewise, galectin 4, whose expression is believed to be restricted to the small intestine, colon and rectum (we observed it in colon and pancreatic tissue) has been shown to be expressed 1.5 to 50-fold lower in colon adenocarcinomas than in adjacent normal mucosa [95]. This value is also in broad agreement with our observed expression change of x5.1. The large number of ribosomal proteins observed as differentially expressed in normal and colorectal tumours, is also in agreement with observations made by Kasai et al. [96] (recent publications have also indicated that defects in ribosomal proteins may cause cancer in Zebrafish [97]). A number of the other genes that were shown to be differentially expressed had previously been shown to be associated with the onset of various cancer in humans. For example both *EGR1* and *NDPKB*, whose expression were shown to be significantly lower in colon tumours than in normal colonic tissue, are putative tumour suppressors [98, 99]. Likewise *HMGIY*, whose expression was not seen at all in normal colon, is a *c-MYC* target and potential oncogene that has been shown to be associated with nuclear factor kappa-B [100]. Further interesting genes include trefoil factor 3, the trefoil family of proteins has been strongly associated with tumour progression [101], and ubiquitin which regulates many key cancer associated genes through ubiquitination[102].

To ensure that the grouping together of different types of tumour samples obtained in different ways (bulk preparations, microdissections etc) was not adversely affecting our results, we analysed a subset of colon libraries. Only libraries that were derived from bulk preparations of adenocarcinomas were compared to bulk preparations of normal tissue. As would be expected fewer genes were detected as expressed in the colon by this smaller set (4617), however only four out of eighty genes that had a p value less than 0.0001[2] in this investigation did not also have a p value less than this threshold in the larger analysis. Consequently, despite these p values being uncorrected, it appears the grouping together of libraries may only lead to the detection of more genes and a potential increased sensitivity in our analysis.

Of the 16814 genes seen expressed in normal or cancerous tissue from one of

[2]A p value cutoff of 0.0001 was used in these analyses as this corresponded to a q value of approximately 0.05

our ten tissues of interest, 191 were exclusively observed in the colon. Of these, only four had a p value less than 0.0001 for differential expression in tumours and could therefore be potential colon-specific drug targets (all had elevated expression in normal tissue):

- ENSG00000186382
- ENSG00000174992 Zymogen granule protein 16.
- ENSG00000183026
- ENSG00000016490 Calcium activated chloride channel 1 precursor.

Literature analysis confirmed that *CLCA1* is only believed to be expressed in the small intestine and colon mucosa [103]. However, in order to undergo a more comprehensive analysis of colon specific expression all EST libraries would have to be included. The two genes without annotation have since been retired from Ensembl, likely due to the mistranslation of the corresponding transcript.

In order to try and validate our results we looked for congruence between our EST expression data and that derived using SAGE. In order to do this we obtained the data pertaining to four SAGE libraries derived from normal and cancerous colon tissue. The mRNA content of these tissue samples was initially represented by 55209 different SAGE tags, however, having removed those tags that did not occur more than once in either normal or cancerous tissue and those that either did not map to an Ensembl gene or mapped to more than one, only 4128 remained. As mentioned above these SAGE tags derived from the work of Zhang et al.[84] who found that 86% of transcripts were represented by less than five tags across all libraries examined, by taking only a subset of these libraries we had likely increased this percentage. Likewise our use of Ensembl genes likely increased the percentage of unmapped tags as Zhang et al. found that only approximately 10% of tags mapped to known Genbank mRNA entries. As only 2839 of these genes were also represented in our EST dataset we could only validate a fraction of our results. To identify any agreement between these sets of data we examined the EST p values for those genes that had scored a p value less than 0.0001 using the SAGE technique. Of the 39 genes whose expression was elevated in normal tissue via SAGE at this threshold, 12 (31%) were also deemed to have elevated expression in normal tissue using the EST technique (using the same threshold). The remaining 27 genes had EST counts that

were not significantly different (with none of the genes showing a significantly higher cancer EST count). When only the direction of expression change was examined (i.e. was the p value greater or less than 0.5) then there was 74% agreement. Examination of 200 genes not deemed significant by SAGE (100 genes either side of the gene with a p value closest to 0.5) found that in 99.5% of cases both techniques agreed that there was no significant difference in expression (one gene was deemed to have significantly greater expression in normal tissue by the EST technique). Although this data argues for at least partial congruence between SAGE and EST data, examination of those genes displaying elevated expression in cancer by SAGE provides confusion. In this set of genes both techniques agreed that ten of them (33%) were differentially expressed. However nine of these were believed to have elevated expression in normal tissue by the EST technique (i.e. the opposite of the SAGE results). Listed below are those genes deemed significantly differentially expressed in both the SAGE and EST analysis. *EEF-1B* was the only gene to display elevated expression in cancer ($q < 0.05$), the rest showed elevated expression in normal tissue in both datasets.

- ENSG00000186676 Elongation factor 1-gamma (*EEF-1B GAMMA*).
- ENSG00000090920 FC fragment of IGG binding protein.
- ENSG00000166165 Creatine kinase, B chain. (*B-CK*).
- ENSG00000130654 Alpha 2 globin. (two tags)
- ENSG00000092841 Myosin light chain 1 (*MLC1SA*).
- ENSG00000167996 Ferritin heavy chain.
- ENSG00000131981 Galectin-3.
- ENSG00000171747 Galectin-4.
- ENSG00000162896 Poly-IG receptor (*PIGR*) .
- ENSG00000143377 Calpactin I light chain.
- ENSG00000161280 Hemoglobin gamma-a and gamma-g chains.
- ENSG00000128016 Tristetraproline (*TTP*).
- ENSG00000125148 Metallothionein-II (*MT-II*).

Since we were unable to find a suitable, publicly available colon microarray dataset, out of interest, we also examined the results obtained for the prostate tissue and the congruence observed between microarray and EST data. Initial investigation of those genes identified as significantly differentially expressed in prostate tumours through the analysis of EST libraries revealed that there was enrichment in this set of genes for those with a previous strong association with cancer. For example, 1 in 9 (17/156) had previously been shown by the Annotation Consortium [104] to have such an association as opposed to the value of 1 in 20 (1106/22184) for the genome as a whole (χ^2, $p<0.001$). Further investigation of this set revealed that it contained potential proto-oncogenes (e.g. rac-alpha serine/threonine kinase [105]) and tumour suppressors (e.g. DAN [106]).

In order to try and validate these results we compared them to those derived using microarrays. Analysis of those Ensembl genes that were deemed significantly overexpressed in normal tissue ($p<0.0001$) showed that where there was corresponding microarray data available for that gene it agreed with the direction of potential expression change (i.e. whether the normal intensity was greater than the tumour intensity) around 72-80% (first 64 or 30 matches) of the time. However, as with the SAGE comparison, when those genes that were believed to be significantly up in cancer were examined there was little agreement (43% over 21 matches).

Therefore of the 81 genes deemed differentially expressed in colon tumours by the EST technique, over a quarter (22) were also deemed differentially expressed in colon tumours using the SAGE technique. This is substantially more than we would expect by chance (hypergeometric cumulative $p < 0.0001$) and supports the hypothesis of congruence between expression platforms. Consequently irrespective of the platform used, a large number of the same candidates will be retrieved in expression studies. However almost half (9) of the genes deemed differentially expressed across both platforms differed in their direction of change. It may be that many genes involved in oncogenesis need only be dysregulated to promote tumour development and there are some known examples supporting this hypothesis. For example both increases and decreases in expression of *MIC-1* have been associated with tumourigenesis [107]. However despite the significantly high congruence between the SAGE and EST analyses a large number of genes were also deemed differentially expressed on only one platform. Whether these genes are less important candidates than those shared across platforms will require further analysis.

2.3.1 Further analysis

A number of genes, such as *tp53*, have been associated with a variety of cancers and consequently we identified those genes that were differentially expressed across tissues. Although 1376 genes were deemed to be differentially expressed in at least one of the six tissues (colon, kidney, liver, lung, prostate and testis) listed in the corresponding appendix table[3], only 55 were deemed differentially expressed in at least four. These did however include previously known general cancer drug targets such as *HSP90* [108].

Analysis of the genomic location of differentially expressed genes illustrated that particular chromosomal locations were enriched for genes likely to be up in tumour tissue (as well as vice versa). This data suggests that the likelihood of a gene being called differentially expressed in cancer may partially be dependent on its chromosomal location. More detailed examination of one of these regions (that mapped to band q13.2 of chromosome 19) highlighted the fact that it contained five genes deemed significantly differentially expressed in colon tumours (Galectin-4, 40S ribosomal protein S16, 40S ribosomal protein S19, FC fragment of IGG binding protein and tristetrapoline) and that three of these were in the list of 13 that were also deemed differentially expressed by the SAGE technique. In total this region contained 89 genes, four of which had been associated with cancer by the annotation consortium (fibrillarin, suppressor of ty 5 homologue, rac-beta serine/threonine kinase and carcinoembryonic antigen CGM2) as well as NF-KAPPAB inhibitor beta. The probability of 5 of 89 randomly chosen genes being deemed significantly underexpressed in tumours is less than 0.0001 (hypergeometric cumulative p). These data could therefore potentially be used in conjunction with those derived by CGH (comparative genomic hybridisation) to identify regions of potential chromosomal aberrations in cancer. There is however a possible alternative explanation for at least some of this non-random distribution, and it is discussed in the next section of this report.

Further analysis of the data allowed us to identify several genes potentially coregulated with genes already associated with cancer (i.e. displaying similar expression profiles across tissues). In an attempt to validate these results we looked for genes whose best correlated partner was a known functionally associated gene (identified by calculating the pearsons coefficient between the expression profiles across tissues

[3]http://www.hgu.mrc.ac.uk/users/james.prendergast/appendixtable.xls

of all genes, and identfying for each gene its corresponding partner with the highest r). A number were identified including several pairs from critical pathways in cancer. These included FASL receptor and *MAPKKK14*; *MAPKKK3* and *MAPKK1*; and *MAPKKKK4* and *NFRκB* which are all associated with the MAPK signaling pathway and were all identified as being coexpressed.

The main aim of this project was to begin to identify candidate genes that may be involved in colorectal cancer. Although this project is a start for such prioritisation it is unlikely to be sufficient by itself. This is because the fact that a gene is differentially expressed in cancer is not by itself enough evidence to say that it is the cause of cancer; nor unfortunately is the observation of a gene not being differentially expressed evidence that it is not involved. The reasons behind this are numerous; gene expression is not the same as protein expression; diseases can be caused by all kinds of factors e.g. polymorphisms, inappropriate splicing etc and this technique is not free from errors (like all gene expression techniques). These data however should, when combined with information from other sources, be useful in characterising the causes and progression of cancer [109].

One of the genes deemed differentially expressed in this study was the transcription factor *EGR-1*, whose expression was shown to be dramatically lower in colon and lung tumours than in corresponding normal tissue (q values of $9x10^{-7}$ and $4x10^{-7}$ respectively). *EGR-1* has been shown to suppress transformed growth in fibrosarcoma cell lines by directly controlling transforming growth factor-beta-1 (*TGFB1*), and its loss of expression can lead to uncontrolled growth in tumours [110]. Subsequent tests in our lab by Dr Farrington have shown that polymorphisms within this gene show an association with colon cancer. Although *EGR-1* was the only gene tested, the use of gene expression studies in this way is consequently likely to be a viable technique for identifying candidates for gene association studies.

Chapter 3

Tumour Gene Expression and Chromatin Structure

3.1 Introduction

In the previous chapter we illustrated that the genomic distribution of genes differentially expressed in cancer may not be random. In agreement with this, genes whose expression increases in tumours have previously been found to be clustered into particular chromosomal regions [111]. Likewise genomic domains have been identified that contain genes co-expressed in cancer [112]. However, there has been little investigation of how this differential gene expression in diseases correlates with global chromatin structure. Recently, a genome-wide analysis of higher-order chromatin structure showed that there are specific domains of the human genome that are enriched in open chromatin fibres [113]. In this study we consequently investigated the potential relationship between such chromatin fibre structures and changes in gene expression in cancer.

3.1.1 Mammalian chromatin arrangement

The primary role of chromatin fibres is the packaging of the two metre long eukaryotic genomes into $10\mu m$ cell nuclei. This substantial compaction of DNA into chromatin is achieved through its association with various key proteins, and the first level of chromatin organisation in mammals is the nucleosome, that consists of two tight superhelical turns of approximately 147 base pairs of DNA around a pro-

tein octomer. Each protein octomer consists of two copies each of the four positively charged histone proteins H2A, H2B, H3 and H4, and this combined DNA and protein package is termed the nucleosome core particle. Each histone protein has been shown to contain two distinct functional domains, a trihelical histone-fold motif required for the histone-histone and histone-DNA interactions within the nucleosome, and tail domains that are subjected to post-translational modification processes, such as acetylation, methylation and ubiquitination. Further histone proteins, termed linker histones (e.g. H1), bind to these nucleosomes and are associated with a further 20 base pairs of DNA [114, 115].

Although the supercoiling of DNA into these nucleosomes reduces its total length by approximately seven-fold, mammalian chromosomes in the interphase nucleus are condensed in the order of 250-fold, with DNA at metaphase condensed even further to around 10,000 times less than its uncompacted length. The nucleosome is therefore only the first level of compaction in the mammalian nucleus. Electron microscopy has shown that under certain conditions each nucleosome along a contiguous length of DNA is separated by 10 to 100 base pairs, in an arrangement likened to that of beads on a string. However these 10nm nucleosomal arrays have been shown to be further condensed into 30nm fibres and finally into structures greater than 100nm in diameter *in vivo*. Although the precise mechanisms and conformation of these higher order chromatin structures are still unknown, it has been shown that if the linker histones or the core histone tails are removed, condensation beyond nucleosomal arrays can not occur, and it has been proposed that linker histones stabilise tail-mediated nucleosome-nucleosome interactions that are the core of higher order chromatin structures [116, 117].

Despite a precise conformation still being unknown, the relatively rigid conformation of DNA around each nucleosome and the constraints imposed by the 30nm size have allowed a number of models to be proposed for the structure of the 30nm fibres of compacted DNA. These include the twisted-ribbon, solenoid and crossed-linker models. However none have been conclusively shown to be the true structure of 30nm chromatin. It may be the case that a number of different arrangements of 30nm fibres are found in vivo [60].

Beyond the 30nm fibre, a number of experiments have demonstrated the existence of large loops of DNA within the mammalian nucleus that can bring together regions of the genome that are initially several Mb apart [118]. It appears that each of these loops is tethered at its base to the chromosome scaffold, a structure that provides

the backbone and characteristic morphology of mammalian metaphase chromosomes and that consists largely of two proteins; SC-1 and SC-2. The degree of looping along chromosomes is however not constant and regions that stain dark upon treatment with the DNA-binding dye Giesma generally contain smaller, tighter loops. These tightly looped "G-band" regions are comparatively gene poor and contain few ubiquitously expressed housekeeping genes compared to the more diffuse structures located in lightly stained "R-bands". There is some evidence that base composition may play a role in determining these banding patterns as G-bands are, in general, more AT rich than R-bands [60].

The final level of chromatin organisation are chromosome territories. A variety of different chromatin types have been shown to occupy particular regions within the nucleus, with gene-poor, mid-to-late replicating chromatin, for example, located at the nuclear periphery [119]. Whole chromosomes have even been shown to occupy distinct nuclear regions, with the relatively gene dense chromosome 19 adopting a substantially more central position in human lymphocyte nuclei than the gene poor chromosome 18 [120]. Likewise the transcriptional status of genes has been shown to be associated with their position within chromosome territories, and it has been proposed that gene repositioning in the nucleus can modulate expression levels [121]. Between chromosome territories themselves is the interchromatin compartment that contains machinery involved in splicing, replication, transcription and repair, and regions of DNA have been shown to loop into this region when active [117, 121].

3.1.2 Chromatin and gene expression

Although, as just discussed, chromatin provides the mechanism for substantial amounts of DNA to be packaged into relatively small nuclei, chromatin also plays a vital role in controlling DNA replication and gene expression. This is because the very compaction of chromosomes that allows their packaging, inevitably makes the corresponding DNA increasingly inaccessible. Transcription machinery must overcome the obstacles of chromatin structure in order to induce gene expression. For example, it has been shown through micrococcal nuclease digestion experiments, that the spacing of nucleosomes around a gene are radically altered upon its transcription (though nucleosomes may still be associated with the corresponding section of DNA). This is as would be expected as RNA polymerases are likely to require the displacement of nucleosome proteins from DNA upon transcription. Likewise experiments in yeast in

which the histone levels within cells were artificially lowered, led to the de-repression of endogenous genes (though the expression of constitutive genes was unaffected). There is therefore strong evidence of at least some form of passive role of chromatin structure in gene expression; those regions of the genome that are transcriptionally competent (rather than necessarily active) are relatively lacking in nucleosomes [60].

Some regions of the genome have even been shown to be completely lacking in nucleosomes. These regions, that are identified through their extreme sensitivity to digestion by DNaseI (and are therefore termed DNaseI hypersensitive sites or DHS), are generally associated with transcription and replication regulatory motifs such as promoters, enhancers and replication origins. It appears that some of these key regions in the genome are protected from nucleosome packaging. For example, the human β-globin locus contains a cluster of DNaseI hypersensitive sites collectively known as the locus control region or LCR. Transgenic experiments in mice have shown that irrespective of its position of insertion in the mouse genome, if the LCR is intact, the chromatin structure of the human globin genes is opened and they are expressed at the correct developmental stage. However mutations or deletions in the LCR lead to random expression of the genes of the globin locus and no protection from the surrounding chromatin structure [60].

The wrapping of DNA around nucleosomes leads to substantial bending of the double helix, consequently some sequences of DNA can surround nucleosomes more readily than others. For example AT base pairs have been shown to preferentially occupy inside positions in nucleosome core particles and GC base pairs outside ones [122]. This is because the bending of DNA around the nucleosome involves the compression of the minor groove on the inside of the DNA double strand and expansion of the major groove on the outside. As the minor groove of AT bases is compressed more readily, they are generally found more centrally. Although no further simple rules or motifs have been identified that determine nucleosome positioning, randomly generated synthetic sequences of DNA has allowed the identification of those sequences that bend effectively around nucleosomes and those that appear unable to form part of a nucleosome core particle. The sequence of DNA consequently plays an important role in nucleosome positioning [60].

As discussed histones must however have some level of mobility to allow transcription, and it has been shown that nucleosomes bound with a strong positioning sequence are still able to slide along the DNA strand. It appears nucleosomes are able to move approximately 10 base pairs at a time so that those base pairs positioned

on the inside and outside of the DNA sequence remain in these respective positions. There appears to be a degree of flexibility in the precise location of a positioning sequence around a nucleosome, that provides the various positions the nucleosome may adopt. These small 10, 20, 30bp etc shifts in nucleosome positioning appear to be enough to provide initial access to the underlying DNA [123, 124, 125].

Transcription is not however only controlled by the positioning of nucleosomes as nucleosomes themselves are not always sufficient to inactivate the underlying DNA. Likewise the cell requires mechanisms for displacing nucleosomes more than the few base pairs possible through positioning sequences. Consequently a number of protein complexes have evolved that are able to displace nucleosomes completely from DNA. These include the SWI and SNF family of proteins, whose role in nucleosome displacement were identified through their affect on DNaseI cleavage patterns of nucleosomal DNA; upon treatment with SWI and SNF proteins, cleavage patterns more closely resemble those characteristic of free DNA [126]. Proteins that modify nucleosomes, such as histone acetyltransferases that attach acetate groups to the N-terminal tails of histones, have also been shown to reduce nucleosome binding. The complement of histone acetyltransferases are histone deacetylases, and together these proteins control the levels of histone acetylation, and consequently to some extent gene expression, across the genome. Further histone post-translational modifications that have been shown to modulate gene expression, include ubiquitination, that assists in maintaining transcription, and methylation, that appears to work in conjunction with acetylation in activating regions of the genome [127, 128].

Beyond the level of the nucleosome the role of chromatin structure in modulating gene expression becomes less clear. As the higher order structures of chromatin are still poorly understood, their roles in gene expression are even less well characterised. For example, although modifications of the N-terminal domain of histones are believed to impact higher-order chromatin structures, the precise alterations in structure, and mechanisms by which they occur, are unknown. Perhaps the best characterised link between higher order chromatin structure and gene expression is that of heterochromatin. Regions of heterochromatin (i.e. those regions that stain strongly), that are generally thought to be more condensed than the majority of the genome, are also thought to generally be less transcriptionally active. However these are both far from applicable to all regions of heterochromatin, and regions of both relatively open and transcriptionally active heterochromatin have been observed. Transgene experiments have shown however that, in general, heterochromatin is able

to silence adjacent genes. It has been proposed that a repressive chromatin structure may spread from blocks of heterochromatin, or alternatively, heterochromatin occupies areas in the nucleus rich in silencing factors to which nearby genes are drawn (the role of nuclear positioning in gene silencing has already been discussed) [60].

3.1.3 Chromatin and cancer

As discussed in chapter 2 altered patterns of gene expression are a hallmark of cancer, but the mechanisms that bring about such extensive changes in transcription are unclear. There is however increasing awareness that chromatin structure plays an important role in controlling the expression of genes, including those with relevance to cancer. The widespread changes in DNA methylation seen in tumours, the efficacy of inhibitors of chromatin modification as anti-cancer agents, and the involvement of chromatin modifying enzymes in cancer, suggest that chromatin structure, at least at the level of the nucleosome, plays a pivotal role in carcinogenesis [129, 130]. Although the molecular detail of higher-order chromatin structure is not known, it is likely that this is also implicated in altered gene expression, both during normal differentiation and in carcinogenesis. For example, the polycomb complex, which compacts arrays of nucleosomes in vitro [131], is implicated in cancer and is involved in regulating the expression of tumour suppressors such as p16Ink4 [132]. Likewise chromosome territories have been shown to be altered within some tumour cell lines [133]. In this study we have therefore investigated the potential role of chromatin structure in tumour gene expression and attempted to determine whether an association between gene expression change in cancer and underlying chromatin structure can be observed. If so, then this would lend support to the hypothesis that chromatin structure and gene expression changes in cancer are inter-related.

3.2 Methods

3.2.1 Chromatin data

Data on chromatin fibre structure was obtained from the Bickmore lab. Their technique for determining chromatin structure is as follows.

Nuclei from lymphoblastoid cells are first digested with micrococcal nuclease, which specifically targets and cleaves DNA between nucleosomes. By using spe-

cific digestive conditions DNA fragments of approximately 10-30kb in length are then obtained. These fragments are then separated on an isokinetic sucrose gradient according to their sedimentation rate. DNA fragments of the same length and chromatin structure will be observed to sediment together, along with shorter fragments with an open chromatin structure as well as longer fragments with more closed DNA. This is because a given length of DNA will sediment faster if it is packaged into a more compact chromatin structure leading to a decrease in its frictional coefficient. To subsequently separate fragments by their size, and consequently also by their chromatin conformation, each sedimentation fraction is run through agarose gel electrophoresis. Fragments that run more slowly than the bulk of the fraction will represent long fragments of relatively closed DNA with the faster moving fragments on the other hand having a more open conformation. In order to analyse the distribution of open chromatin across the genome, differentially labeled open and input chromatin fractions are cohybridised to a genomic DNA microarray. In this analysis the array used was assembled from clones spaced at ~1Mb intervals, from the 'golden path' used in the sequencing of the human genome (16). The data used per clone was the average of four hybridisations, performed with colour reversal. Regions of the human genome enriched with open chromatin will display a $\log_2$ open:input ratio that is greater than 0 [113].

3.2.2 Expression data

81 adult, human EST Expressed Sequence Tag (EST) libraries were obtained from dbEST (`http://www.ncbi.nlm.nih.gov/dbEST`) that were non-normalised, non-subtracted and not obtained using the ORESTES [85] technique. They each contained a minimum of 250 sequences and were grouped according to tissue (colon, kidney, liver, lung, prostate or testis) and disease state (normal or cancer). To identify the gene of origin for each EST all 449,365 ESTs were RepeatMasked (A.F.A. Smith and P. Green, unpublished results) and subsequently compared to the Ensembl cDNA database (version 18) using BLAST [87]. Any EST that matched an Ensembl cDNA sequence with 95% identity over at least 100 base pairs was assigned to that cDNA. In this way approximately 66% of the ESTs were successfully assigned to a known gene with the remaining unmapped ESTs a result of unknown transcripts or poor sequencing (as mapping success varied dramatically between EST libraries, even those of the same tissue, it is likely a large proportion of these ESTs are a result

of poor library preparation).

13 adult human Serial Analysis of Gene Expression (SAGE) [82] libraries derived from colon, lung or prostate tissue (normal or tumour) were obtained from CGAP (`http://cgap.nci.nih.gov/SAGE`). Tag counts were grouped according to tissue and disease state, and mapped to genes via data available at the NCBI website (`ftp://ftp.ncbi.nlm.nih.gov/pub/sage`). Two colon adenoma libraries (derived from Familial Adenomatous Polyposis patients) were also obtained from SAGENET (`www.sagenet.org`) and compared to the normal colon libraries using the same technique.

Three microarray datasets comparing normal and tumour tissue were used in this analysis [134, 135, 136] (see table 3.1). The Affymetrix CEL data files of each microarray dataset were analysed using the dChip software [137] (`http://www.dchip.org/`). If appropriate information was available, tumour samples with less than 40% tumour cells were removed before analysis. To exclude arrays with potential image contamination or sample hybridisation problems any chip with an array outlier percentage greater than 5% was also discarded [138]. Using data available at the Affymetrix website (`http://www.dchip.org/`) probes were subsequently assigned to genes.

Author	Array type	Tumour type	Normal arrays	Tumour arrays
Bhattacharjee et al. 2001^{+}	U95Av2	Lung adenocarcinoma	17 (16)	127 (72)
Lenburg et al. 2003*	U133A	Renal cell carcinoma	8 (8)	9 (9)
Singh et al. 2002 $^{+}$	U95Av2	Prostate tumour	50 (49)	52 (52)

Table 3.1: Microarray datasets used in this analysis.
Numbers in brackets indicate the number of arrays left after the quality control stage and consequently used in the analysis. $^{+}$Dataset available at: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. *Dataset available at: http://www.ncbi.nlm.nih.gov/geo/

Gene coexpression was measured through the use of the GNF Gene Atlas V2 Human U133A microarray collection [139]. The 158 raw CEL image files for each array of this dataset (79 tissues, 2 replicates of each) were analysed using the dChip software [137]. Using dChip each array was first normalised at probe intensity level to the array with median overall intensity and then expression values were calculated using the PM/MM (perfect match/mismatch) [137] difference model (truncated values to 1).

29 outlier arrays were removed in which the model-based standard error of greater

than 5% of probes was greater than three times the median of all PM-MM probe pairs (as implemented in dChip) and the remaining expression values were subsequently log transformed. Redundant probes were masked (by only keeping the first occurring probeset of those that mapped to the same LocusLink id, as in dChip) and all probes that could not be mapped to a unique position in the genome (NCBI build 35) with at least 90% identity were excluded (arguably a higher threshold could have been set but we allowed for both mismatches as well as polymorphisms).

Expression values of replicate arrays were pooled so that each gene was represented by 68 values, one for each tissue type remaining after the above filtering (often only one of the two arrays for a tissue were lost after filtering and consequently only 12 tissues were completely unrepresented). Those genes that were potentially tissue specific (i.e. those with a present call in less than 20% of the arrays, the default cutoff in dChip) were also filtered out as we were not interested in those genes that were purely expressed in the same single tissue but those that showed some level of coregulation. As a match at a single tissue would lead to a large Pearson's r we excluded these genes with restricted expression profiles.

To minimise the affect of large values the expression values across all tissues for each gene were standardised to have a mean of 0 and a standard deviation of 1. Pearson's correlation coefficients were subsequently calculated between these expression values of all genes in the same chromatin category, and the proportion of comparisons above a chosen threshold calculated.

3.3 Results and Discussion

During the investigation of tumour gene expression via the analysis of EST libraries discussed in the previous chapter, we identified regions of the genome enriched with differentially expressed genes. However further analysis illustrated that even whole chromosomes differed in their average p values, and that some chromosomes contained a greater proportion of genes that were up-regulated in cancer than others. As nuclear localisation is believed to play a role in gene expression, we investigated whether an association between a chromosomes position in the nucleus and its average gene expression change could be observed. As shown in figure 3.1 a simple analysis seemed to suggest that this was indeed the case. However as nuclear localisation is also correlated with gene density which itself is correlated with chromatin

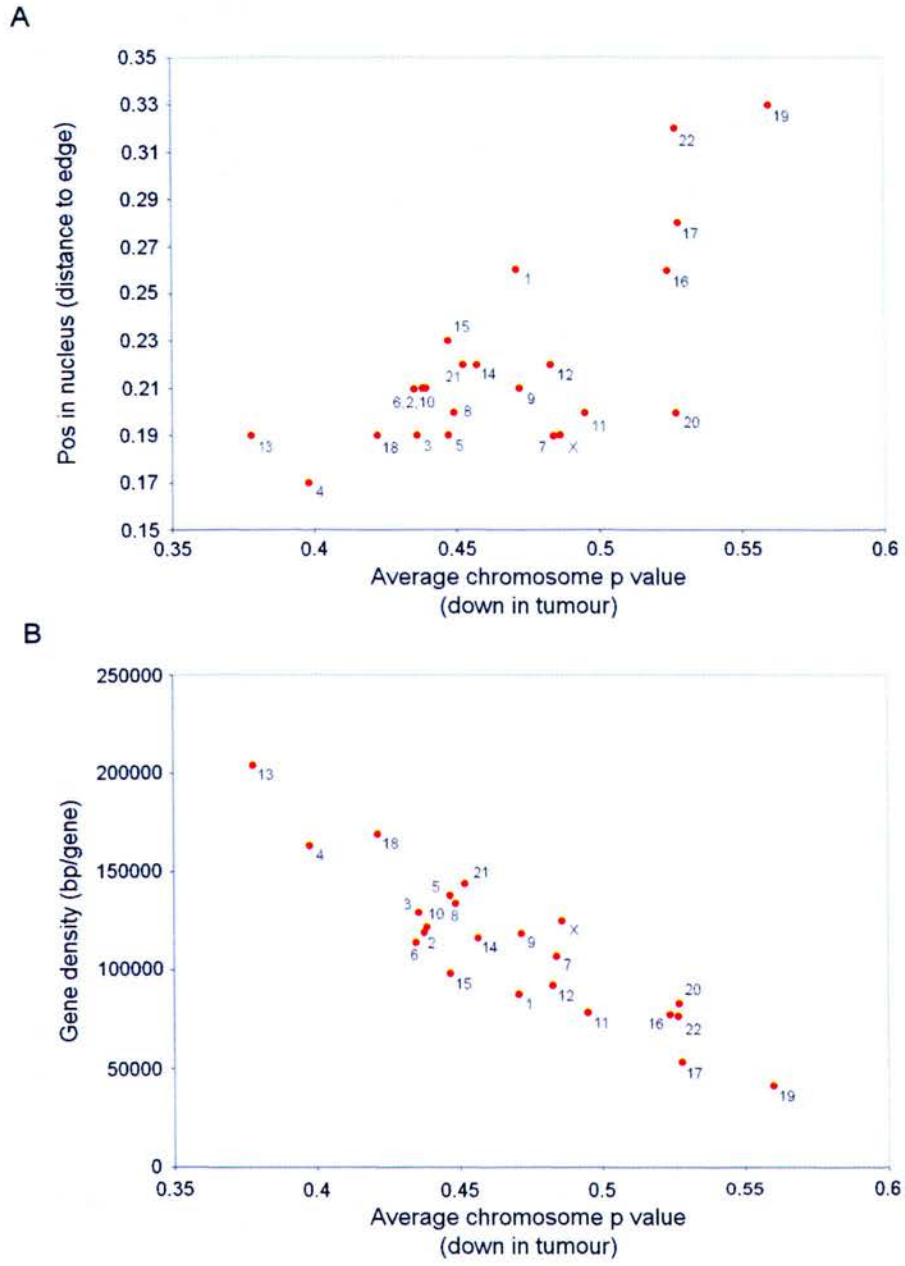


Figure 3.1: Gene expression change and nuclear localisation
The relationship between gene differential expression in cancer and a chromosome's (A) position in the nucleus and (B) average gene density

structure we investigated whether a correlation between gene expression and chromatin structure could be observed *within* chromosomes.

In 2003 Zhou et al. used EST sequences to investigate tumour gene expression. The analysis of Zhou et al. [111] identified chromosomal domains in which the genes are generally up-regulated in tumours compared to normal tissues. They termed these, regions of increased tumour expression (RITEs). We noted, in a comparison to the Gilbert et al. dataset, that the distribution of RITEs both along, and between, chromosomes was qualitatively similar to that of domains of structurally open chromatin fibres [113]. For example, on chromosomes 1 and 11 the largest RITEs appear to align with the regions that are highly enriched in open chromatin fibres at positions; 0-45Mb on 1p34-36, 144-153Mb on 1q21, 0-20Mb on 11p15, and 63-76Mb at 11q13 (Figures 3.2+3.3). It was also noted that RITEs are very sparse on chromosome 13, whereas a very large proportion of the genes on chromosome 19 are up regulated in tumours. Similarly, chromosome 13 is rather depleted of open chromatin fibre domains whereas most of chromosome 19 is highly enriched for open chromatin fibres (Figure 3.3).

We believe however that the analysis of Zhou et al. suffers from a number of drawbacks. The first, is that Zhou et al. used normalised and subtracted EST libraries. Normalisation and subtraction of EST libraries increases the sampling of rare transcripts by preferentially forming duplexes of abundant sequences. Although these techniques have proven useful in gene discovery, they degrade the quantitative relationship between transcripts. The extent of normalisation and subtraction also differs across libraries and consequently expression levels based on these EST collections are not representative of the true RNA abundance in a cell.

The technique of Zhou et al. also generally leads to gene dense regions of chromosomes being scored higher than those regions with a more sparse coverage of genes. This is due to the scanning index used by Zhou et al. To determine clusters of genes up-regulated in tumours, they first calculated the TMZ (trimmed mean of Z-score) for 500kb regions every 100kb along the genome. Regions of the genome with consecutive windows of high TMZ were then scored more highly depending on the number of consecutive windows above a cutoff. Gene poor regions however are simply less likely to contain consecutive windows, and as the presence of open chromatin fibres across the human genome has been shown to correlate with gene density, this apparent correlation between chromatin and gene expression change may not actually be the result of the upregulation of genes in open chromatin in cancer.

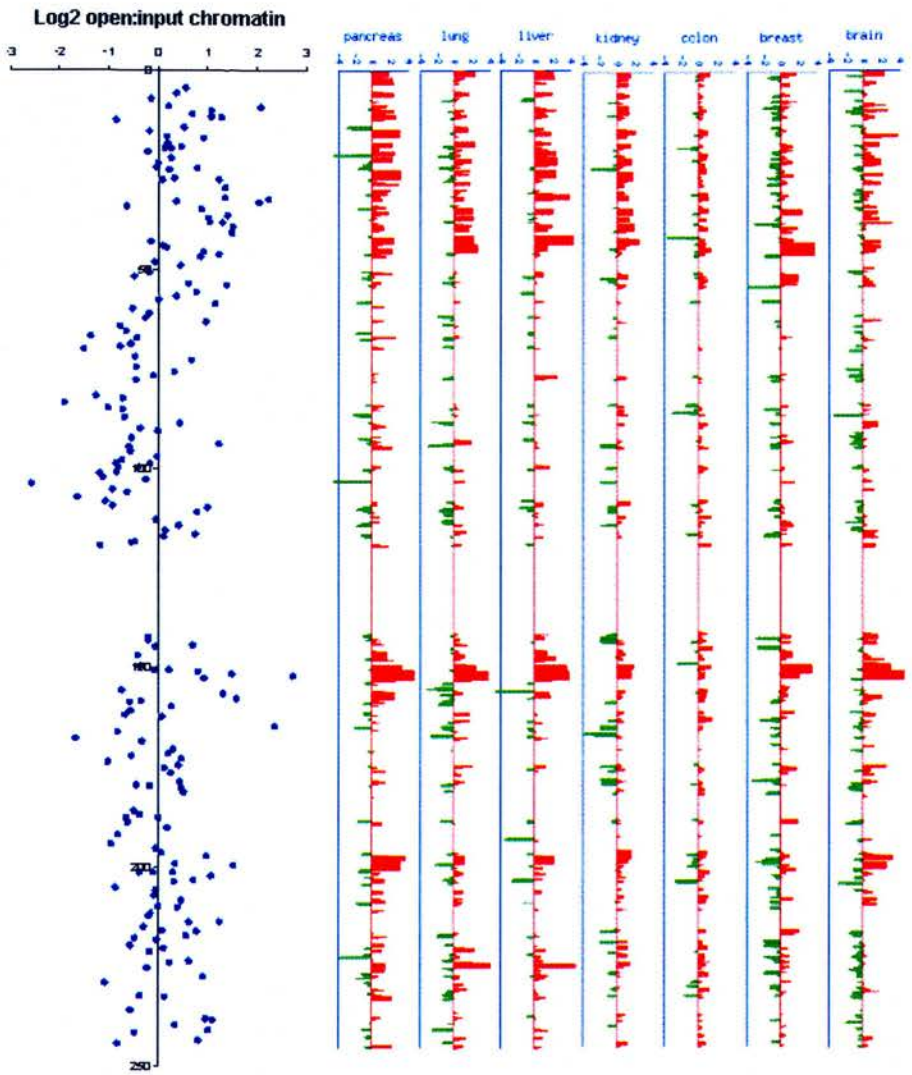


Figure 3.2: Comparison between the gene expression analysis of Zhou et al. and chromatin structure (I).
The $\log_2$ ratio of open:input chromatin fibres prepared from lymphoblastoid cells, is shown for the BACs from the 1Mb array from human chromosome 1. Aligned to the right, is the gene expression analysis of Zhou et al. [111] for the same chromosomes. Regions of increased tumour expression in pancreas, lung, liver, kidney, colon, breast and brain tumours are shown in red. Regions of decreased tumour expression are shown in green.

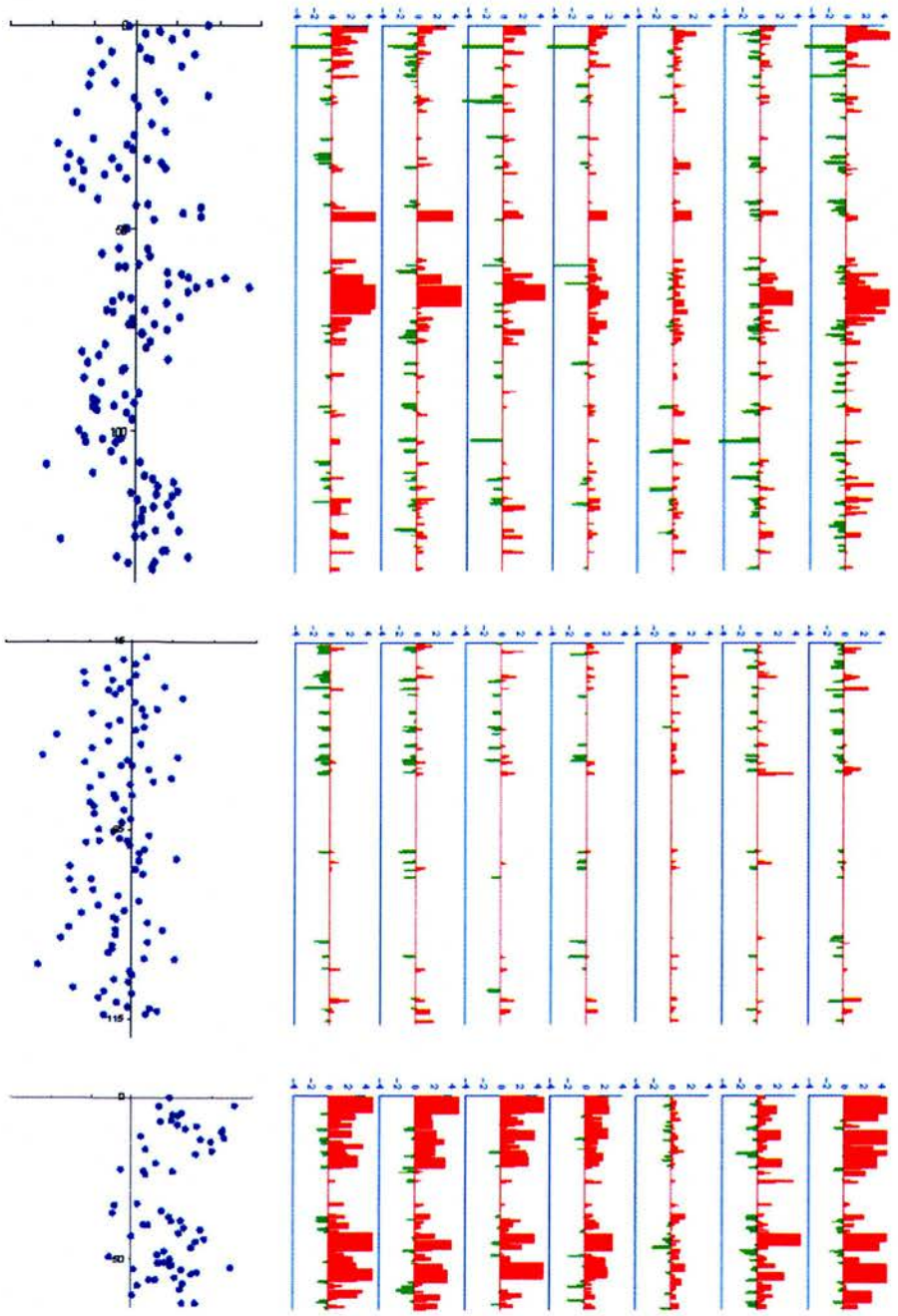


Figure 3.3: Comparison between the gene expression analysis of Zhou et al. and chromatin structure (II).
The $\log_2$ ratio of open:input chromatin fibres prepared from lymphoblastoid cells, is shown for the BACs from the 1Mb array from human chromosomes 11, 13 and 19. Aligned to the right is the gene expression analysis of Zhou et al. [111] for the same chromosomes.

To determine whether a relationship between chromatin structure and expression change in cancer does in fact exist, as the data of Zhou et al. suggests, we adopted a more simplified technique for analysing gene expression across chromosomes. Unlike Zhou et al., we had no need to determine which regions of the genome were *significantly* differentially expressed, and we therefore simply calculated the average $\log_2$ gene expression change in sliding windows across the genome (using the EST expression data from chapter 2). Each sliding window was 500kb in size and all windows containing less than three genes of known gene expression were excluded (as an average fold change based on only one or two genes may be dramatically skewed from the true mean expression change of that region). As shown in figure 3.4 peaks of gene expression change appear to coincide with areas enriched with open chromatin. These initial results consequently led us to test for a relationship between chromatin and gene expression change directly.

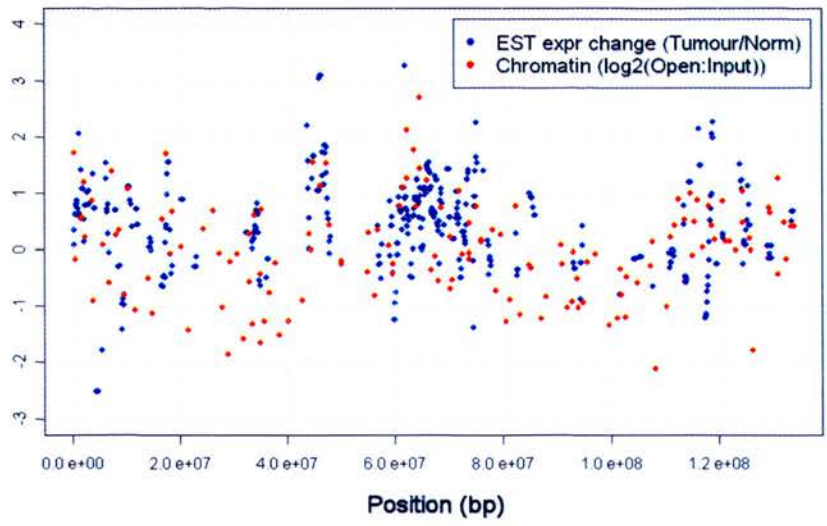


Figure 3.4: The chromatin structure and gene expression change in cancer observed across chromosome 11.
Clone $\log_2$(open:input) values are represented in red and window average gene expression change in blue.

Adopting this sliding window approach had two major problems. First it suffered from gene density effects and second it did not allow us to directly compare chromatin structure and gene expression change, as at best the sliding windows only partially overlapped regions of known chromatin structure (and therefore quantification of the

relationship between chromatin and gene expression change was difficult). To more precisely compare the changes in gene expression in cancer with the chromatin fibre structure of the human genome, we collected EST data for six normal tissues (colon, kidney, liver, lung, prostate or testis) and for cancers that involve these tissues. In each case, the relative ($\log_2$) change in expression observed for each gene in cancer, relative to the corresponding normal tissue, was calculated using algorithm 1.

Algorithm 1 Calculation of gene expression change

$$GeneExpressionChange = log_2(\frac{NumCancESTsAssigToGene/TotalNumCancESTs}{NumNormESTsAssigToGene/TotalNumNormESTs})$$

Genes were excluded if corresponding ESTs could not be found in either normal or tumour libraries, or if they were represented by less than 4 ESTs in total. The average $\log_2$ expression change was then calculated across the six tissues. Finally the average $\log_2$ fold change in expression of all the genes that fell entirely within a BAC clone from the chromatin 1Mb genomic array was calculated. This was termed the clone average $\log_2$ fold change (CALFC). All BAC clones with a CALFC value > 0 were said to be up regulated in cancer. A significance cutoff was not applied as we were looking for a trend across all clones rather than just in those deemed to be significantly differentially expressed. BAC clones to which no genes mapped were excluded from analysis. We grouped BAC clones according to their $\log_2$ open:input chromatin value. BACs with the most open chromatin fibre structure were grouped as those clones with $\log_2$ open:input < 2.5 and >1.5. Clones with $\log_2$ open:input of between 0.5 - 1.5 were the next group, followed by those between -0.5 and 0.5. Regions with the most closed chromatin fibre structure were grouped together as those BAC clones with $\log_2$ open:input $<$ -0.5 and $>$ -1.5. Within each of these groups, the proportion of BAC clones for which the CALFC value was > 0 was then determined. As shown in figure 3.5 there is a clear correlation between CALFC values and chromatin structure. A large proportion (70%) of BAC clones with a very open chromatin structure ($\log_2$ open:input >1.5) have a CALFC value > 0. Conversely, only 40% of BAC clones from regions of the human genome depleted of open chromatin ($\log_2$ open:input -1.5 to -0.5) have CALFC values > 0.

We were however keen to ensure that a relationship between gene expression change and chromatin structure could also be observed using other expression platforms. The correlation between chromatin fibre structure and gene up-regulation in cancer was therefore also examined for the SAGE and microarray expression plat-

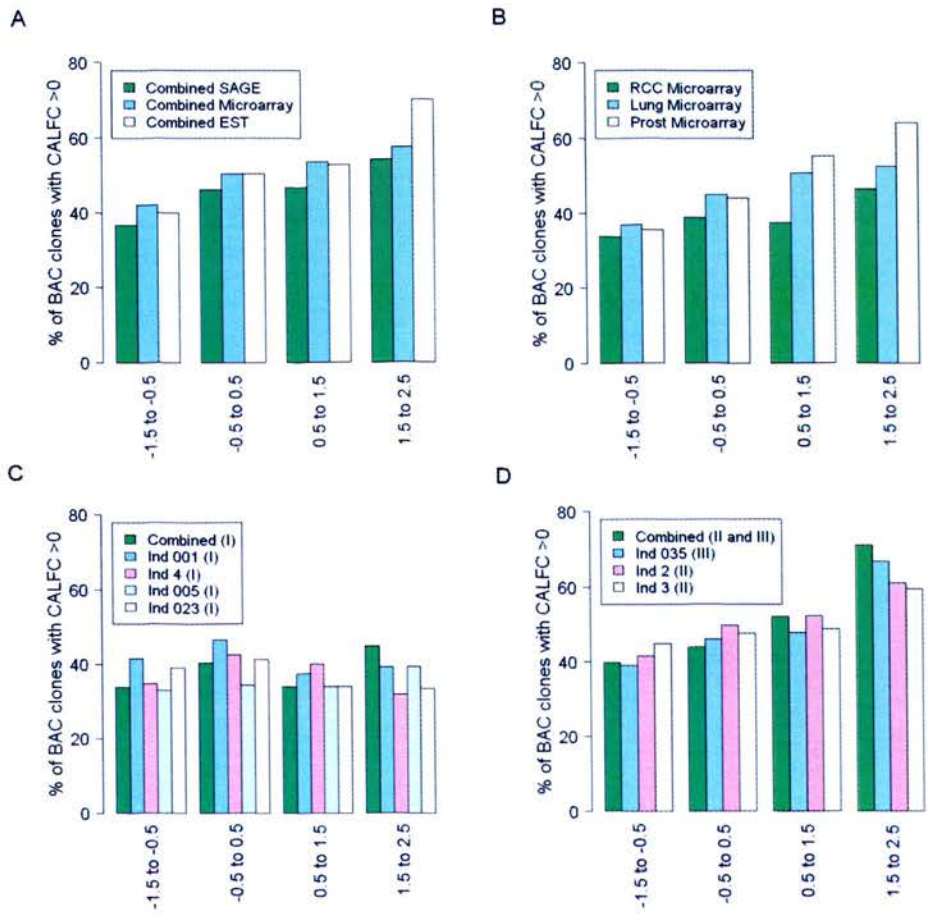


Figure 3.5: The correlation between chromatin fibre structure and CALFC values (A) Histograms showing the percentage of BAC clones from the 1Mb array whose average $\log_2$ fold change in expression in cancer (CALFC) was > 0. BAC clones were grouped according to the $\log_2$ open:input chromatin fibre structure hybridisation ratio. Clones with the most open chromatin fibre structure are those where $\log_2$ open:input >1.5. Clones with the most closed/compact chromatin fibre structure are those where $\log_2$ open:input $<$ -0.5. Data from three different expression platforms; EST (open bars), SAGE (shaded bars) and microarray (filled bars) datasets, combined from six different tissues, is compared. (B) Analysis as in A, but comparing microarray data from three different tissue/tumour types; kidney/RCC (open bars), lung (shaded bars) and prostate (filled bars). (C) Histograms showing the percentage of BAC clones from the 1Mb array whose average $\log_2$ fold change in expression in cancer (CALFC) was > 0, for RCC tumours at Fuhrman stage 1. (Individual ids are those from the original Lenburg et al. paper [134], numbers in brackets following each id indicate the individuals Fuhrman stage) (D) Histograms showing the percentage of BAC clones from the 1Mb array whose average $\log_2$ fold change in expression in cancer (CALFC) was > 0, for RCC tumours at Fuhrman stages 2 and 3.

forms. $\log_2$ fold gene expression changes of SAGE tags or microarray data were calculated as for ESTs (only the U95Av2 datasets were used in the combined microarray plot). As with EST data, it was found that the regions of the human genome most enriched in open chromatin fibres had the highest proportion of BAC clones with CALFC values >0 (Fig. 3.5A) (combined EST: r^2=0.78, p=0.004; SAGE: r^2=0.476, p=0.058; Microarray: r^2=0.655; p=0.015). Therefore we conclude that genes are most readily activated, or up-regulated, in cancer if they derive from regions of the human genome that have an open chromatin fibre structure.

The graph in Figure 3.5A is the result of combining expression data derived from six different normal/cancer tissue types. None of these tissues corresponds to that used to investigate chromatin fibre structure - lymphocytes - however chromatin structure is not believed to change dramatically between most cell types (personal communication - Professor Wendy Bickmore). To examine whether the correlations between CALFC and chromatin fibre structure are cell type dependent, data from individual tissue types was examined (Fig. 3.5B). A significant correlation between CALFC and chromatin fibre structure was found in each case (prostate: r^2=0.862, p=0.001; lung: r^2=0.546, p=0.036; RCC: r^2=0.689, p=0.011).

Tumour grading by pathological analysis of nuclear morphology and morphometry indicates that very gross changes in chromatin/nuclear structure correlate with later stages of tumourigenesis, and with generally poor prognosis and survival. For example, Fuhrman nuclear grading for renal call carcinoma (RCC) is a good predictor of disease survival [140]. Grade 1 tumours have small round, evenly stained nuclei. Nuclei at grade 2 are more irregular in shape and staining, with mildly enlarged nucleoli, and this is more prominent in grade 3 tumours. At grade 4, cells are very enlarged and pleiomorphic [141]. We subdivided the RCC microarray dataset according to Fuhrman grading of the analysed tumours (as defined in [134]), and examined the correlation between chromatin fibre structure and average gene expression change for tumour grades 1 to 3 (Fig. 3.5C+D). We observed no significant correlation in grade 1 tumours (p=0.11) but a strong correlation in those defined as grade 2 and 3 (p<0.001), with the grade 3 tumour displaying the greatest change between open and closed categories (67 to 39%). A similar correlation to cancer progression was seen for SAGE data in colon cancer. Differential gene expression in colon adenomas (benign tumours) does not display the same relationship with chromatin structure (r^2 = 0.034, p=0.82), as is seen for adenocarcinomas (r^2 = 0.98, p<0.01) Figure 3.6.

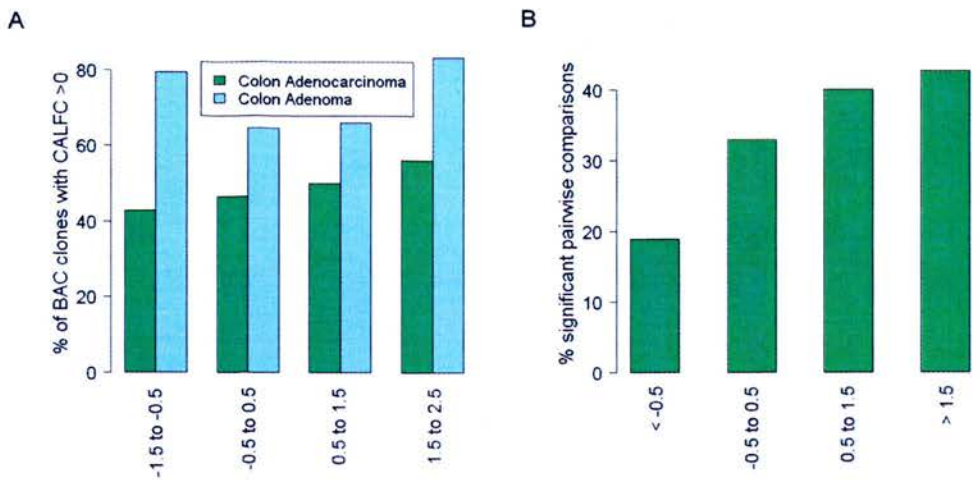


Figure 3.6: Chromatin structure, staging and coexpression
(A) Histogram showing the percentage of BAC clones from the 1Mb array (16) whose average $\log_2$ fold change in expression in cancer (CALFC) was > 0, for SAGE data obtained from two colon adenoma libraries and from two colon adenocarcinomas. (B) Comparison of the levels of coexpression among genes in different chromatin categories. The Pearson's correlation coefficient was calculated between the expression profiles of all genes in the same chromatin category and the proportion of pairwise comparisons with an r greater than 0.24 is shown (A cutoff of 0.24 was used as each expression profile contained 67 values, one for each tissue examined, and with 66 degrees of freedom, the significance threshold of r at an α of 0.05 is 0.24). $r^2$of proportions versus chromatin as shown in panel B =0.90, p=0.05.

We had one major concern about this analysis however. Traditional methods of analysing microarray data are dependent on the assumption of constant variance across all levels of gene expression and often that the data are normally distributed. However in 2001, Rocke and Durbin showed that microarray data are more accurately modelled by the equation shown in algorithm 2. From this equation it can be observed that when expression levels are large the middle error term dominates and the raw (measured) expression value is approximately equal to μe^{η}. However when expression levels are near background, i.e. μ is close to 0, then the middle error term is almost insignificant and the measured expression level is approximately equal to simply $\alpha+\varepsilon$. Consequently, as $log(y) = log(\mu)+\eta$, simple log transformations should sufficiently stabilise the variance of expression data at high levels. However the closer expression levels get to background levels the less applicable a log transformation will be, and the more inflated the variance of observations. As variance will also not be symmetrically distributed around the true expression level, measured expression levels of genes of low expression will be artificially skewed.

Algorithm 2 Rocke and Durbin's two-component model of microarray error.

y the measured raw intensity, α is the mean background noise, μ the true expression level and η and ε are normally-distributed error terms with mean 0 and variance σ_{η}^2 and σ_{ε}^2, respectively.

$$y = \alpha + \mu e^{\eta} + \varepsilon$$

Gilbert et al. [113] had observed no correlation between gene expression and chromatin structure, and therefore this poor performance of traditional log transformation was initially of little concern. However if certain chromatin categories were enriched with genes of low expression, and the skewed variance was not identical in both normal and tumour samples, this could lead to artificial correlations between gene expression and chromatin structure. We therefore retested the hypothesis that chromatin structure and gene expression were independent. To do this we obtained the gene expression profiles of normal human B cells from the study of Klein et al. [142] (as chromatin structure was determined in lymphoblastoid cells). Comparisons of clone average gene expression values to the $\log_2$(open:input) scores of the corresponding clones, suggested, that there was in fact a relationship between gene expression and chromatin structure. As can be seen in figure 3.8A the average

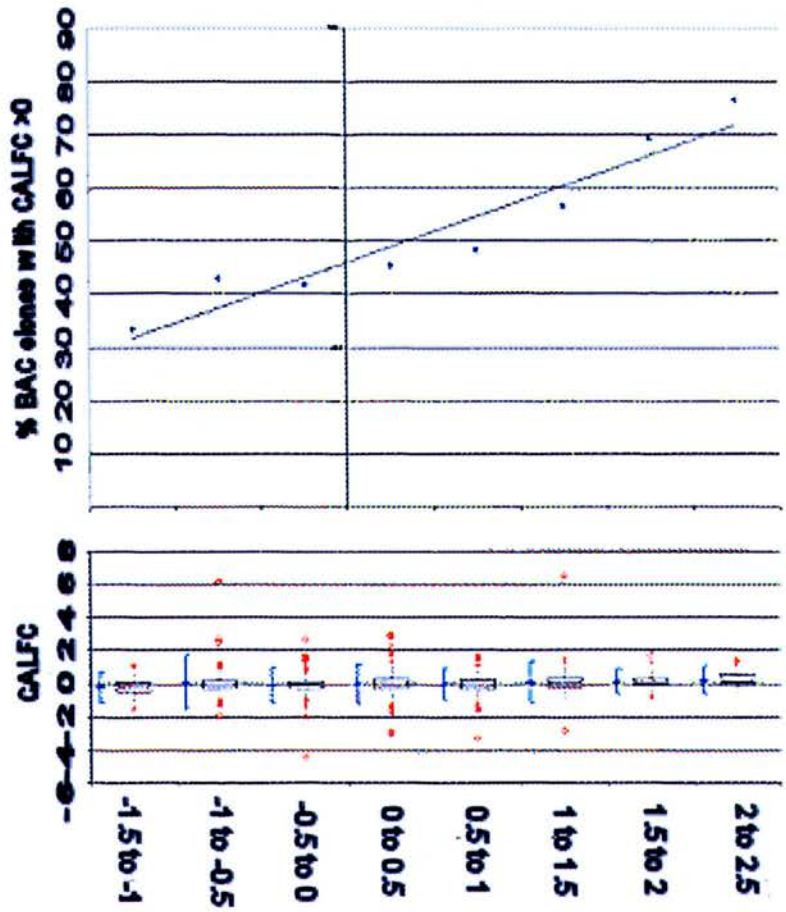


Figure 3.7: Combined RCC stage II and III results shown in Figure 3.5, split into smaller categories (-1.5 to -1, -1 to -0.5 etc) with corresponding boxplot.
The blue diamonds of the boxplot mark the means and corresponding confidence intervals of the CALFC values in each category. The notched boxes show the medians, their confidence intervals and the lower and upper quartiles. The dotted lines connect the nearest observations within 1.5 inter-quartile ranges of the lower and upper quartiles. Red crosses and circles mark near (greater than 1.5 inter-quartile ranges away) and far (greater than 3 IQRs) clone outliers respectively. This figure illustrates that although there are relatively large changes in the proportion of BAC clones with a CALFC value > 0 for each chromatin fibre structure category, the actual difference in the mean CALFC between clones in each category are relatively small. The linear relationship between the chromatin structure and CALFC values of clones was also confirmed by linear regression and it was shown to be weakly but significantly correlated (r^2=0.011, p=0.001)

expression of clones with a more open chromatin structure is generally higher than in more closed clones.

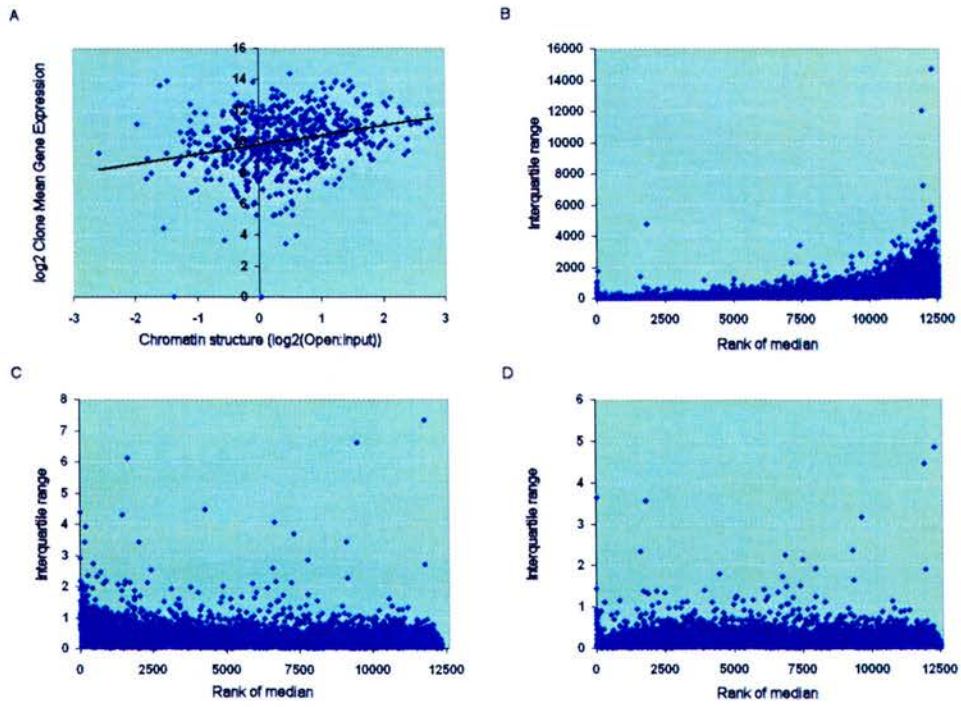


Figure 3.8: Hawkin's transformation
(A) Clone chromatin structure versus log transformed average gene expression (r2 = 0.066, p = $5x10^{-11}$). (B) The rank of the measured intensity versus the IQR for raw probe intensities. (C) The rank of the measured intensity versus the IQR for log transformed probe intensities. (D) The rank of the measured intensity versus the IQR for Hawkins transformed probe intensities.

To test therefore whether variance between replicates was indeed affected by a probes expression level we ranked the probes of the same Klein B cell dataset according to their median intensities. As shown in figure 3.8B the probes with medians near 0 displayed relatively similar variances in intensities. However, those probes with large median intensities showed large variations between repeats. Although, as shown in figure 3.8C, log transformation of the data can lead to a stabilisation of the variance of probes with large expression intensities, a large number of probes with low medians are outliers and display a high variance.

These results therefore agree with the two-component model of microarray data predicted by Durbin and Rocke, and in conjunction with the apparent correlation

between gene expression and chromatin structure, highlight a potential issue with our analysis. This is because the variance observed in these microarray datasets can lead to skewed values of genes with low expression profiles. Randomisations carried out by Durbin et al. have shown that lowly expressed genes can show relatively large negative skews in their values upon simple log transformation. If therefore this underestimation of the expression of probes near background is not identical in both the normal and cancer array sets, an artificial correlation between gene differential expression and chromatin structure may result. We consequently applied the variance-stabilising transformation of Hawkins to try and negate the relationship between expression level and inter-replicate variance. This stabilisation is a solution of the algorithm shown in algorithm 2 and is itself shown in algorithm 3. Application of this transformation, as illustrated in figure 3.8D, does lead to stabilisation of the variance observed across expression levels and reanalysis of our data using this transformation does not lead to the loss of the correlations observed between gene expression change in cancer and chromatin structure (Figure 3.9). We therefore hypothesise that the correlations observed above are not the indirect result of a correlation between gene expression and chromatin structure. This is however based on the assumption that the transformation of the normal and cancer datasets are perfectly matched (even Durbin et al. have illustrated that this transformation does not completely negate the skewing of genes of low expression). As this transformation is also designed to stabilise the variance between experimental repeats, and not as in our datasets inter-person replicates, further analysis will be required to confirm without question that gene expression levels are not playing a role in creating the observed correlations. This will require the creation of a novel gene expression study that can be designed from the start to be controlled for the affects of gene expression levels across chromatin structure.

Algorithm 3 Hawkins microarray data transformation [143].

c, a constant determined from the data, equals $\sigma_\varepsilon^2/\sigma_\eta^2$.

$$z = ln[(y-\alpha) + \sqrt{(y-\alpha)^2 + c}]$$

We are confident however that the correlations we have observed are real as there are other levels of evidence that suggest that the relationship between gene expression and chromatin structure is unlikely to be leading to the correlation between gene expression change and chromatin structure. For example, we only ever observed

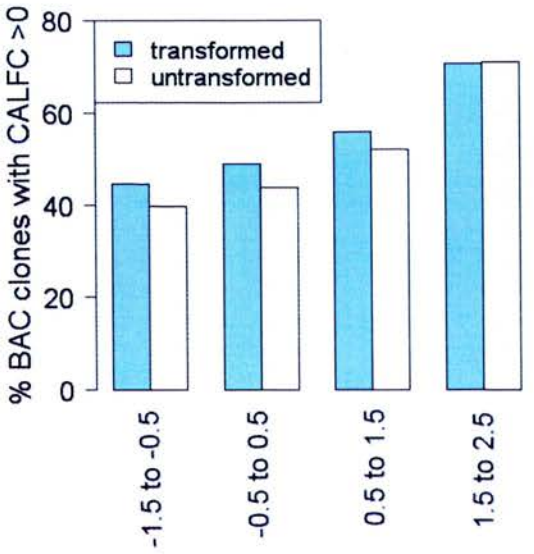


Figure 3.9: RCC stage II and III results with and without variance stabilising transformation having been applied.
With the transformation the results marginally improve (With transformation $r^2 = 0.96$, without transformation $r^2 = 0.95$).

positive correlations. There is no reason to believe that normal microarrays will be consistently more (or less) skewed than cancer arrays. Likewise the observation of changes in the strength of the relationship between gene expression change and chromatin structure with tumour stage does not seem to be consistent with what we would expect if skewed data was the cause of the correlations. Only if the relationship between gene expression and chromatin structure was affected by disease state or stage could skewing lead to these correlations, and although this may be expected, we could find no evidence that this is the case.

So why do the majority of regions of open chromatin display a mean increase in gene expression in later stages of certain cancers (Figure 3.5)? Does the expression of genes in open chromatin generally change in the same direction across tissues? In an attempt to answer this question we used the Novartis Gene Atlas V2 human expression dataset [139] to determine whether there was a higher level of coexpression between genes in open chromatin than those in closed, i.e. do genes from open chromatin regions generally show more similar expression profiles to one other than those in closed chromatin? This was indeed found to be the case. As shown in

figure 3.6 when the expression profiles of all genes in open chromatin were compared to one another, 43% of the pairwise comparisons were above the significance cutoff of an r of 0.24. This is compared to only 19% in closed chromatin. Although the tissues used in this analysis were not independent and consequently the significance cutoff used in this analysis may be too lenient, the same results were also observed with higher r^2 values (a weighted Pearson's would have been more appropriate). Likewise we observed a correlation between chromatin structure and the average r^2 value between genes, i.e. the more open the chromatin category the higher the mean r^2 was when the expression profiles of all the genes within that chromatin category were compared.

Figures 3.10 and 3.11 illustrates that there are at least two distinct expression patterns in open chromatin; those genes highly expressed in tissue from the brain and those genes highly expressed in tumours and various other tissues. A potential cluster of genes highly expressed in the brain is also seen in closed chromatin but further clearly defined clusters are less easily observed. Only four genes in closed chromatin are strongly expressed in tumours, and of these the two that show strong expression in more than one cancer type (*MAD2L1* and *EIF4E*) have already been strongly associated with cancer [144, 145]. It appears therefore that chromatin may have a key role to play in controlling the expression of genes, particularly those in open chromatin, and that this role may be dysregulated in cancer.

We have therefore found a correlation between gene expression changes in cancer and the recently described chromatin fibre architecture of the human genome. Our analysis provides a potential mechanistic basis for the observed genomic clustering of genes over-expressed in cancer that had previously been reported [111]. The regions of the human genome where genes are generally upregulated in cancer correspond to those regions that have been shown to have a biophysically open chromatin fibre structure [113]. This correlation was found for three different expression platforms; ESTs, SAGE tags and microarray datasets. This is particularly notable given the rather modest congruence that we, and others, had previously observed between these platforms [146]. The presence of open chromatin fibres across the human genome has been shown to correlate with gene density, but not with gene activation or silencing per se [113]. It has been suggested that these domains of open chromatin fibre structure provide a constitutively open environment that facilitates transcription. Indeed, genes that are widely expressed in normal tissues tend to cluster in these domains [113, 147, 148]. Consistent with this idea, we found a corre-

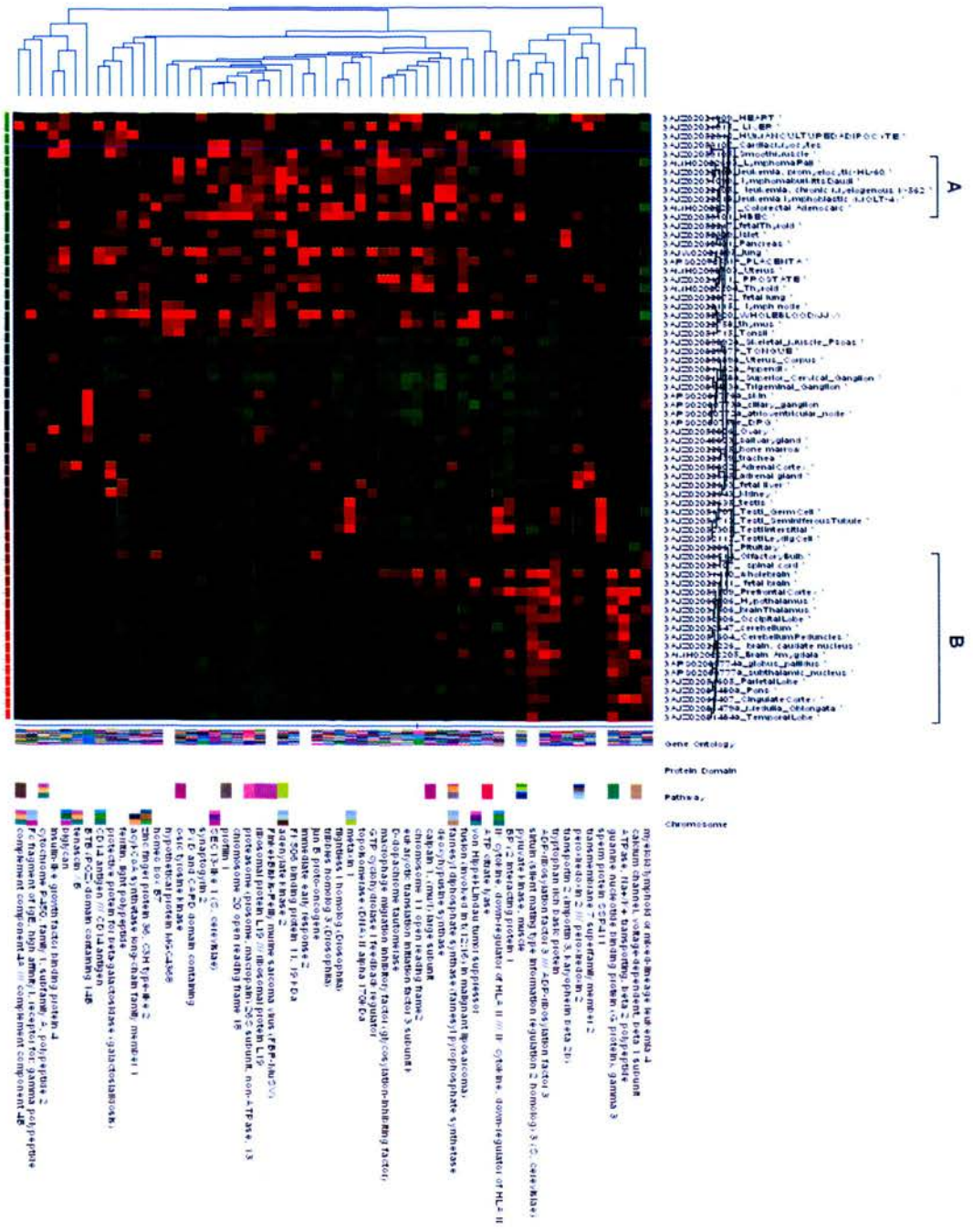


Figure 3.10: Cluster analysis of genes located on the most open clones. Genes were clustered according to their standardised expression in 68 tissue types, with the distance between genes calculated as 1 - r (where r is the pearsons coefficient). Expression values were clustered by both gene and tissue. For more details see `www.dchip.org` A: Cluster of arrays derived from tumour tissue, B: Arrays derived from brain and spinal cord tissue.

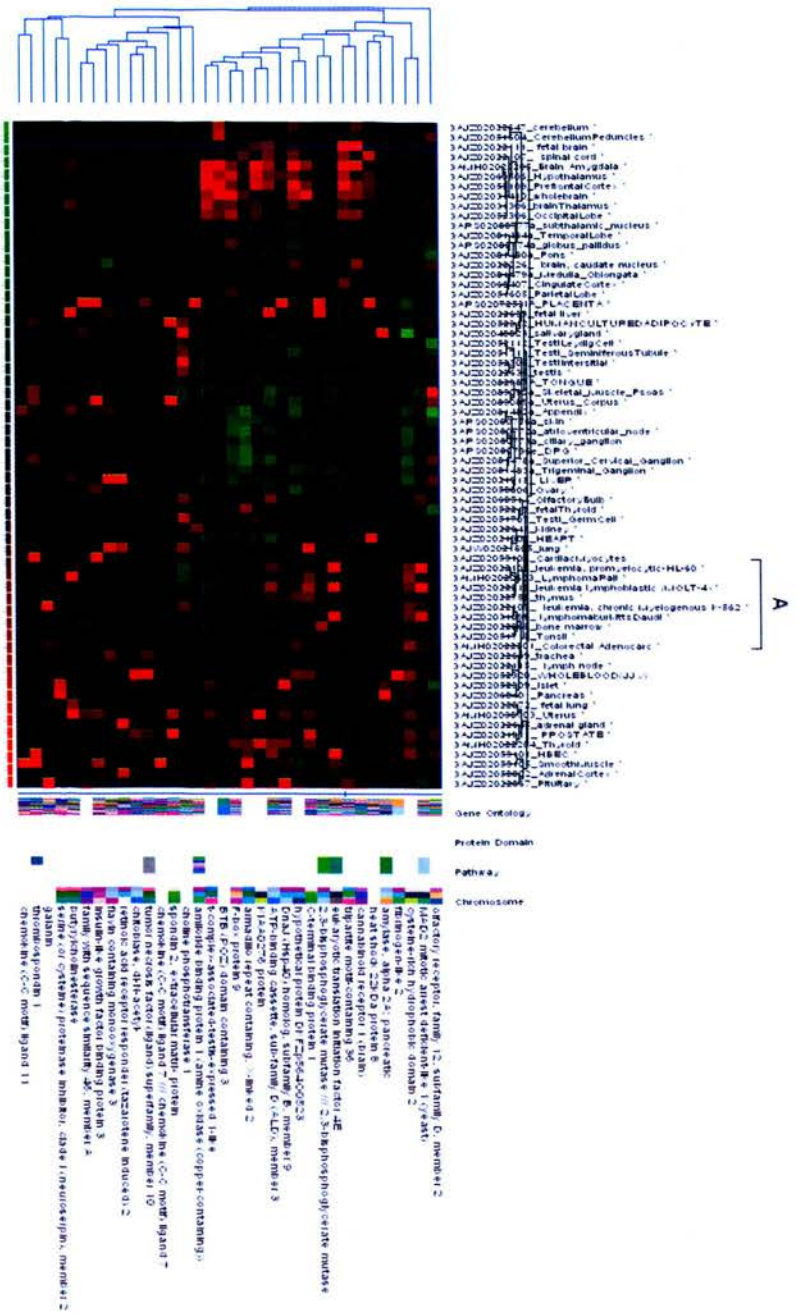


Figure 3.11: Cluster analysis of genes located on the most closed clones. Genes were clustered according to their standardised expression in 68 tissue types. A: Cluster of arrays derived from tumour tissue. Note that cluster A is interspersed with arrays derived from thymus, bone marrow, and tonsil.

lation between transcriptional up-regulation and these same domains in a variety of tumours that are derived from different tissue-types. We also observed higher levels of gene coexpression between genes in open chromatin than those in closed regions of the genome. However, the magnitude of changes in gene expression in cancer is quite modest, clones within the most open chromatin fibre domains show on average only a 16% increase in expression in stage II and III renal cell carcinomas. This could be because the basal level of gene expression of genes within these regions is already quite high in normal tissues [148]. In contrast, expression of genes within domains that have a generally closed/compact chromatin structure may require remodeling of the higher-order chromatin structure before they can be activated during disease. The correlation of gene-expression changes and chromatin fibre structure appears to be most prominent for the later stages of cancer progression. This suggests that the structure of chromatin itself may be altered during cancer progression. Understanding how and why this occurs may provide new insight into the mechanistic basis of gene-expression changes and cancer progression.

Chapter 4

Chromatin Structure, Mutation and Selection

4.1 Introduction

As discussed, regions of open and closed chromatin structure have recently been defined across the human genome [113]. In chapter 3 we highlighted the relationship between this chromatin structure and gene expression change in cancer. However the mechanism behind this observed relationship remains unclear. Why do genes in open chromatin generally show an increase in gene expression in later stages of tumourigenesis? Although we illustrated that genes in open chromatin generally show high levels of coexpression, this does not provide an explanation for the initial driving force behind this correlated change in gene expression. However one potential mechanism behind this phenomenon is mutation. DNA mutation is one of the key factors underlying oncogenesis and loss of efficient DNA repair is a hallmark of most if not all cancers. Although many mutations associated with tumourigenesis have been shown to affect a protein's structure, many have also been shown to modulate a gene's expression levels, for example through changes at a gene's regulatory region. It is plausible that any given mutation at a regulatory motif in open chromatin is more likely to lead to an increase in gene expression than a reduction. This is because an open chromatin structure provides the most conducive possible environment for transcription, consequently regulatory motifs in open chromatin are more likely to be limiting a gene's expression than those in closed regions. Single base mutations may also explain the relatively small changes in expression observed in our analysis,

copy number changes or large scale insertions or deletions would likely lead to larger differences in expression levels between normal and tumour tissue.

Mutation rates may also underlie the paradox surrounding chromatin structure and gene expression. If open chromatin fibre domains provide a chromatin environment more conducive to transcriptional activation, why are genes also found in regions of closed chromatin structure if this simply means they are less accessible for transcription This question can not simply be answered by the compaction requirements of DNA as the vast majority of the genome is non-coding. One possibility is that some genes need to be subject to especially tight transcriptional regulation, and that their aberrant or leaky expression in inappropriate cells cannot be tolerated. However, it has also been proposed that open chromatin structure may make the underlying DNA sequence more susceptible to DNA damage [149]. Consequently certain genes are located in closed regions of the genome due to the low mutation rates they confer.

Although some studies have predicted that rates of mutation are relatively constant across mammalian genomes, analysis of human-mouse alignments has suggested that there may be as much as a 3-fold difference in substitution rates across chromosomes [150], with regions containing genes involved in extracellular communication displaying unusually high levels of synonymous substitutions [151]. Previous studies have also shown that, in mammals, genes within close genomic proximity undergo similar rates of mutation and evolution [152, 153, 151]. For example, Williams and Hurst showed that the mean difference between the Ka values (substitution rate at non-synonymous sites) of 176 pairs of linked genes was significantly lower than would be expected by chance [152]. Similar results were also observed with Ks (substitution rate at synonymous sites) and Ka/Ks (that is often used to infer the mode and strength of selection). Consequently they proposed that the murid genome was split into domains of evolution. Why this was the case was unknown, but it is possible that some aspect of chromatin structure over different genomic regions influences the rate of DNA damage or its repair. Similar analyses in to the clustering of genes with similar expression profiles has shown that a much smaller proportion of human and mouse genes (3-5%) show evidence of restricted gene order [154]. This analysis consequently argues for only limited constraint on the clustering of coexpressed genes.

The availability of a map of long-range chromatin structure across the human genome [113] allows us to assess this idea. In this study we have investigated the

rates of selection and mutation across chromatin categories. By excluding CpG sites we have exclusively looked at mutation rates not attributable to the inherent mutability of methylated cytosines. This is obviously important as open chromatin is enriched with CpG islands. To fully investigate the potential relationship between mutation rates and chromatin structure we looked at various measures of neutral mutation; including intronic, intergenic and ancient repeat divergence as well as SNP density. We also looked for any link between chromatin structure and selection in the human genome.

4.2 Methods

The abundance of open chromatin fibre structure in lymphoblastoid cells, at clones spaced approximately 1Mb apart along the human genome was determined as previously described. Relative chromatin structure is represented in this analysis by $\log_2$(open chromatin:input chromatin) values; determined by cohybridising differentially labeled "open" and input chromatin fragments to a human genomic DNA microarray. A large $\log_2$(open:input) value in this analysis indicates a region enriched with open chromatin. See above or Gilbert et al. [113] for further details. The 2,787 human genes that mapped to each of these clones and their corresponding mouse orthologues were obtained from Ensembl. Coding sequence alignments of each of these orthologous pairs were derived via protein alignments (using the MUSCLE [155, 156] and tranalign [157] programs) and the PAML [158] package was used to calculate dN (rate of divergence at nonsynonymous sites), dS (rate of divergence at synonymous sites) and dN/dS. Gene pairs with anomalously high dS values (>1.270 i.e. twice the median dS of all Ensembl human vs mouse pairs) were excluded.

Rates of transitions in alignments were estimated by observing the positions in each alignment where a transition appeared to have occurred i.e. an A was aligned with a G or a T was aligned with a C. At all other positions in the alignment where bases differed between mouse and human, the mouse base was made the same as the human base and values of dN and dS were recalculated. An analogous process was used to estimate the rate of transversions. It should be noted that this technique will miss those transitions that have been masked by a second change at the same site, so that it appears only a transversion has occurred (and vice versa). However as this analysis was carried out in human-chimpanzee comparisons multiple

substitutions at the same site should be extremely rare. Human vs chimp orthologues were also obtained from Ensembl and again they were filtered according to their dS values. However alignments with very low identity ($< 85\%$) were also excluded. Due to the limited number of differences between human and chimp orthologues, all alignments of genes in the same chromatin categories (i.e. in regions of similar chromatin structure) were concatenated and rates of dS, dStransi and dStransv were calculated for these combined alignments.

The posterior probability that a gene harbours excess amino acid variation (P-) as calculated by Bustamante et al. was obtained from [1] . These data are based on the assumption that purifying selection will lead to a greater proportion of non-synonymous polymorphisms in a gene, relative to synonymous polymorphisms, when compared to the proportion of fixed non-synonymous differences (relative to synonymous differences) found between human and chimpanzee. The hypothesis being tested is that the product of the selection coefficient and population size, γ ($\gamma = 2N_e s$), is not different from 0. Genes are annotated as under positive or negative selection via quantifying the posterior probability that its selection coefficient is less than or greater than 0 given the respective observed data for that gene. If the equal tail credibility intervals of a gene's selection coefficient fall entirely below 0 the gene is deemed to be under purifying selection. See Bustamante et al. for more details [159]. In our analysis we used those genes with at least two variable non-synonymous sites, i.e. Bustamante et al.'s INS (Informative only about Negative Selection) dataset. Of these 6,033 genes, 690 mapped to one of our clones of known chromatin structure. For each chromatin category we calculated the proportion of these genes with a high P-, i.e. greater than 0.95.

Human chimpanzee divergence was determined through the use of the chained and netted human-chimpanzee alignments available at the UCSC website (hg17-panTro1) [160]. Ensembl gene predictions were used to identify intronic, intergenic and protein coding regions. All exclusively intergenic and intronic regions found within clones were identified, and divergence measured in the corresponding sections of the human-chimpanzee alignment using PAML's baseml with the REV model. Before calculating divergence all sequence from the same chromatin category was concatenated and all bases that overlapped a CG dinucleotide in either species were removed from the alignments to conservatively calculate non-CpG rates of divergence

[1] http://www.nature.com/nature/journal/v437/n7062/extref/nature04240-s2.txt

[161].

To identify potential genes with CpG islands, the positions of predicted CpG clusters were obtained from the UCSC genome browser [162]. Of these islands, any that were less than 500bp long, had a G+C content less than 55% or had a CpG to expected CpG ratio of less than 0.65 were excluded [163]. Those genes whose 5' end was within 2kb of one of these islands were determined to be potential CpG island genes. Genes annotated as housekeeping by Hsiao et al. [164], i.e. called as present by Affymetrix software in all 19 tissues examined, were obtained from hugeindex.org.

Intergenic repeats were identified through UCSC's RepeatMasker annotation. Ancient repeats were defined as in Gibbs et al. [165] and Taylor et al. [166] as repeats from the same RepeatMasker subfamily conserved between mouse and human in the same orientation. Simple repeats and regions of low complexity were excluded.

The SNP Consortium data (TSC) were used to calculate SNP density across chromatin categories [167]. To ensure these densities were not biased as a result of the variety of protocols used to detect SNPs (some of which were chromosome specific), SNP densities across chromatin categories were also calculated using only SNPs randomly identified via the TSCM0019 protocol (a panel of 24 DNAs sequenced by the Sanger Centre, for more details see `http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewTable.cgi?method_id=581`). By using this one protocol, that covered the whole genome and not just certain chromosomes, any bias resulting from the variety of different protocols used should have been removed. The location of TSC SNPs was determined by mapping their ssIds to current rsIds via data available at dbSNP.

Predicted Exonic Splice Enhancers (ESE) hexamers were obtained from Fairbrother et al. [168]. The occurrence of each of these hexamers in the coding regions of each of the genes that mapped to a 1Mb clone was determined. In order to identify the number of hexamers we would expect to detect by chance given the base composition of the genes and hexamers, we randomly shuffled the bases in each of the coding regions 100 times and recalculated the occurrence of each of the hexamers. The distribution of non-protein coding genes across chromatin categories was determined through Ensembl annotations.

The proportion of SNP pairs on each clone that displayed strong evidence for recombination or strong LD (linkage disequilibrium) were calculated using the method of Gabriel et al. [169] (using the Haploview program [57]). Clones less than 100kb in length or with an average SNP density less than one every 5kb were excluded. Clones

were subsequently split into categories of similar chromatin structure ($\log_2$(open:input) of -2 to -1.5, -1.5 to -1 etc) and the average proportion of pairwise comparisons displaying evidence for recombination or LD was calculated for each group. This was done for all four HapMap populations [170]. To determine whether regions of relatively closed chromatin display significantly low levels of recombination, the proportion of markers that display strong evidence for recombination and that are less than various distances apart was calculated in the 60 clones with the lowest $\log_2$ open: input values (again clones less than 100kb in length were excluded). This process was repeated with all clones and results compared. Randomisations were also carried out by selecting and analysing in the same way 60 random clones from the 1Mb cloneset, 1,000 times for each population. In all populations the proportion of pairs that were less than 150kb apart and that showed strong evidence for recombination in the 60 open clones was less than the corresponding value in 95% of the randomisations.

4.3 Results and Discussion

4.3.1 Non-dS measures of mutation are highest in closed chromatin

In order to determine whether mutation rates are associated with chromatin structure we first determined intergenic divergence rates, using human versus chimpanzee whole genome alignments, in regions whose chromatin environment in human lymphoblastoid cells had been determined. The majority of intergenic bases should be under little or no selection and therefore intergenic divergence should be approximately analogous to mutation rate. As shown in Figure 4.1, we see a negative correlation between intergenic divergence and chromatin structure. However as open chromatin is generally more gene rich than closed, and may therefore contain more regulatory elements than intergenic regions, we also examined divergence rates in ancient repeats only, but these also displayed the lowest divergence rates when in open chromatin.

It has been proposed that DNA sequences nearer the centre of the nucleus may be protected from DNA damage by those on the periphery (the "bodyguard hypothesis"). Likewise the chromosomes most enriched with open chromatin are generally situated towards the centre of a nucleus [149]. The correlation observed between di-

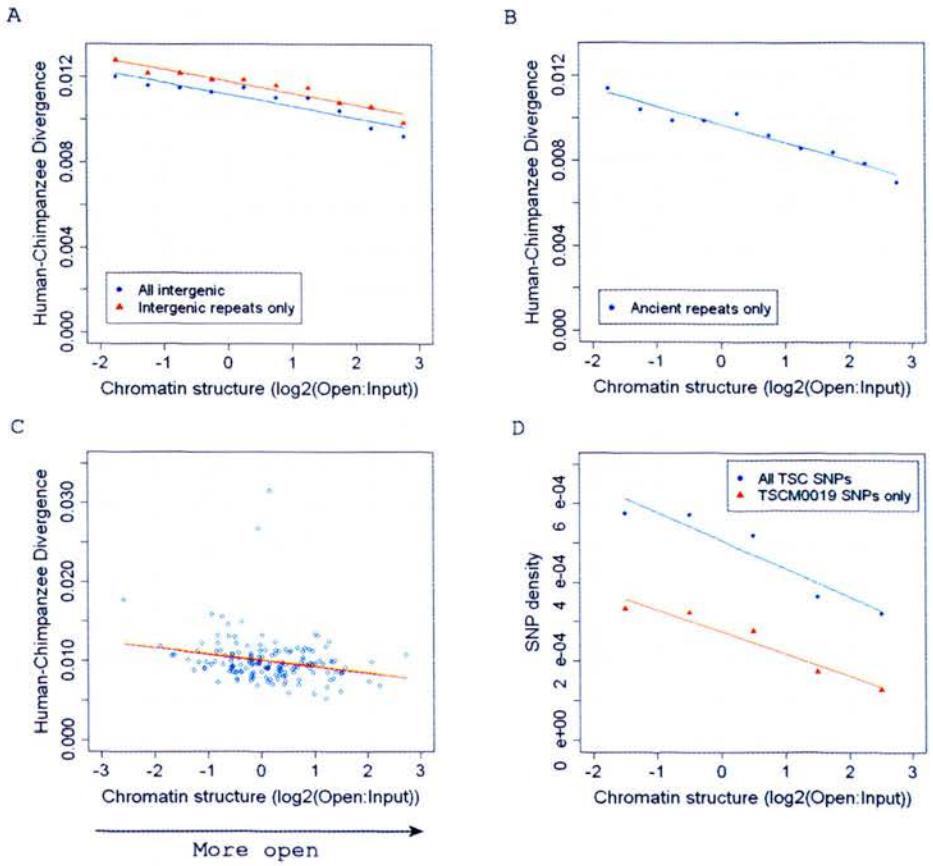


Figure 4.1: Increased mutation rates in closed chromatin.
(A+B) Mean intergenic and ancient repeat divergence observed across chromatin categories(Intergenic: r^2= 0.87, p = 9x10^{-5}; Intergenic repeats only: r^2= 0.93, p = 7x10^{-6}; Ancient repeats only: r^2= 0.93, p = 6x10^{-6}). (C) Intergenic divergence of each 1Mb clone from chromosome 1 against their corresponding chromatin score (10 clones containing less than 10,000 intergenic bases were excluded). (D) Mean human SNP densities (SNPs/bp) observed across chromatin categories (All SNPs: r^2= 0.89, p = 0.016; Single random detection protocol (TSCM0019) SNPs only: r^2= 0.93, p = 0.008).

vergence rates and chromatin structure may therefore be an indirect result of these phenomena. We therefore investigated whether a correlation between intergenic divergence and chromatin structure could be observed within chromosomes. Although chromosomes themselves have been shown to display some level of polar organization such that their most gene-poor regions are those closest to the nuclear periphery [171], adjacent intergenic regions within chromosomes often have very different chromatin structures despite displaying approximately the same nuclear localisation. If the observed correlation between intergenic divergence and chromatin structure reflects the predictions of the bodyguard hypothesis, we would expect to see no such correlation within chromosomes. This however is not the case, for example, as shown in Figure 4.1 there is a negative correlation between intergenic divergence and chromatin structure within chromosome 1 (r^2= 0.053; p = 0.0043). The two outlier clones observed in this figure, with a divergence greater than 0.025, could represent mutational hotspots in the genome. However the degree of difference between the divergence observed in these clones compared with the rest of the chromosome suggests to us that the alignment in these regions are more likely be of poor quality. Removal of these clones increases the significance of the correlation observed between divergence and chromatin structure (r^2= 0.113; p = $3\text{x}10^{-5}$). In total 7 out of 22 chromosomes display a significant negative correlation (p<0.05) between clone intergenic divergence and chromatin structure (Chromosomes 1 p=0.012, 2 p=0.033, 5 p=0.014, 8 p=0.016, 12 p=0.046, 17 p=0.025 and 20 p=0.038; false discovery rate, or FDR, analysis using the q value package suggests at most one of these is expected to be a false positive). These data therefore argue against the bodyguard hypothesis being solely responsible for these observed correlation between chromatin structure and mutation rate.

Another measure often used to predict mutation rate is SNP density [172, 173]. It is predicted that as the majority of intergenic sequence is non-functional and that there has been little time for selection to act on SNPs, their density along the genome should generally reflect underlying mutation rates. A further benefit of the use of SNPs in this way is that mutation rates can be predicted without relying on sequence comparisons with other species. We consequently determined the mean intergenic SNP densities observed across chromatin categories. As shown in Figure 4.1 the mean SNP density was lowest in the most open regions of the genome.

There is therefore strong evidence that mutation rates are associated with chromatin structure. Not only are intergenic, intronic (Figure 4.4) and ancient repeat

divergence rates highest in closed chromatin but the density of SNPs is also elevated in the most closed regions of the human genome and we hypothesise that closed regions of the genome are simply less accessible to DNA repair mechanisms.

Through the use of the DAVID program [174] we identified the classes of genes most over-represented in closed chromatin, and therefore likely to be experiencing the highest mutation rates. Of the 148 genes in the most closed regions of the genome 40 encode glycoproteins (p for enrichment: 0.000074, modified Fishers Exact test) and 22 were associated with the G-protein coupled protein signaling pathway ($p = 0.00011$). As this group of genes contained less than 500 distinct GO terms both these results remained significant after Bonferroni multiple testing correction. Glycoproteins and G-protein coupled receptors are involved in immune response and cell signaling and it has previously been proposed that such genes are likely to evolve quickly in response to changing stimuli [151]. Being located in closed regions of the genome, where we have observed mutation rates are particularly high, will allow this more rapid evolution. Genes that have previously been shown to evolve relatively slowly, i.e. genes such as housekeeping genes, are also preferentially located in open regions of the genome where mutation rates are relatively low (Figure 4.2). The location of a gene in the genome and its subsequent local chromatin structure may therefore at least partly be governed by the suitability of the local mutation rate it confers.

4.3.2 dS is highest in regions of open chromatin

	-2 to -1	-1 to 0	0 to 1	1 to 2	2 to 3	p
dN	0.11	0.12	0.094	0.089	0.075	$2.2\text{x}10^{-6}$
dS	0.62	0.64	0.65	0.66	0.68	0.13
dN/dS	0.18	0.18	0.14	0.13	0.11	$5.8\text{x}10^{-9}$
dN (without CpG island)	0.14	0.14	0.12	0.11	0.082	$3.4\text{x}10^{-4}$
dS (without CpG island)	0.65	0.64	0.68	0.7	0.7	0.045
dN/dS (without CpG island)	0.21	0.22	0.17	0.15	0.12	$4.16\text{x}10^{-6}$
dN (with CpG island)	0.073	0.095	0.071	0.075	0.07	0.043
dS (with CpG island)	0.58	0.64	0.62	0.64	0.66	0.53
dN/dS (with CpG island)	0.13	0.14	0.11	0.11	0.1	0.012

Table 4.1: The raw numbers of the data shown in Figure 4.3 as well as the corresponding p values for the correlation between each data type and chromatin structure when the genes are not binned by chromatin category.

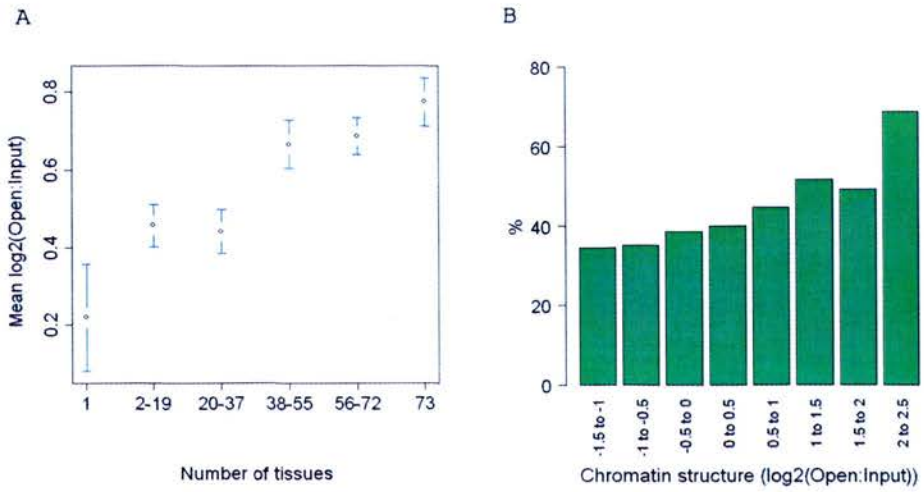


Figure 4.2: The distribution of housekeeping and CpG island genes across chromatin categories.

(A) The mean chromatin structure of genes by expreesion breadth (B) The percentage of genes in each chromatin category that are associated with a CpG island.

dS has historically been used as a further surrogate measure of basal mutation rates, as synonymous sites were believed to be under little or no selection. Changes at synonymous sites, unlike at non-synonymous sites, do not affect the encoded amino acid, and due to the relatively small effective population sizes of mammals, a synonymous site would have to experience relatively strong selection to evolve in a non-neutral manner [175]. As shown in Figure 4.3 the average rate of non-synonymous changes (dN) observed in human mouse alignments is 51% higher in the most closed chromatin regions of the genome than in the most open. Similarly the ratio of non-synonymous to synonymous substitution rates (dN/dS), frequently used as a measure of selection, is 61% higher. However, the average synonymous rate (dS) for genes in relatively open chromatin is higher than that for genes in a more closed chromatin structure. This is consistent with the reported high Ks for human chromosome 19, the human chromosomes with one of the most open chromatin structures of all [176]. The observation by Hurst et al. of similar levels of human-mouse dS, dN and dN/dS in linked genes is likely therefore to be the result of linked genes being from similar chromatin environments. However, to ensure the

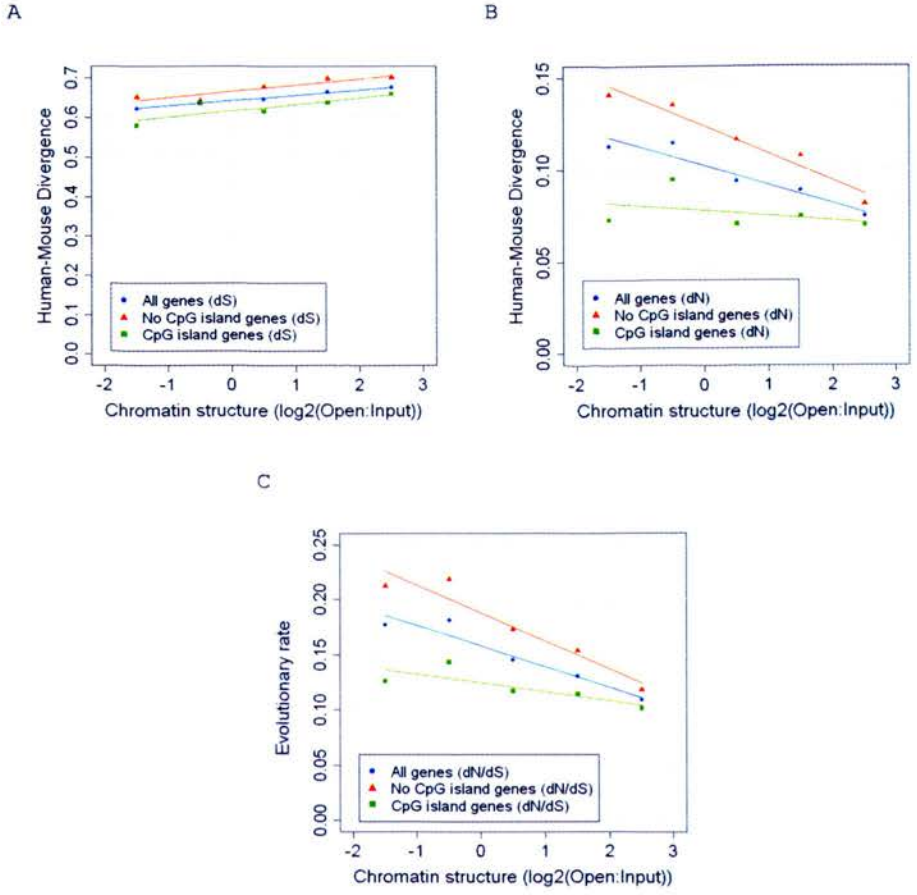


Figure 4.3: Human-mouse divergence observed across chromatin categories. The correlations observed between chromatin structure and mean dS (A), dN (B) and dN/dS (C) in human/mouse coding sequence alignments. (All protein coding genes dS: r^2= 0.99, p = 0.001; dN: r^2= 0.92, p = 0.01; dN/dS: r^2= 0.92, p = 0.009. Genes associated with a CpG island dS: r^2= 0.72, p = 0.07; dN: r^2= 0.17, p = 0.5; dN/dS: r^2= 0.64, p = 0.1. Genes not associated with a CpG island only dS: r^2= 0.84, p = 0.03; dN: r^2= 0.95, p = 0.005; dN/dS: r^2= 0.92, p = 0.01.). If values are not binned then chromatin structure and both dN and dN/dS remain significantly correlated across all genes (p = $2.2\text{x}10^{-6}$ and $5.8\text{x}10^{-9}$ respectively) however the p of the correlation between dS and chromatin structure rises to 0.13.

converse is not true, and that the results observed in this study are not the result of linked genes, we randomly selected only one gene from each clone (so that all genes analysed were approximately 1Mb apart and therefore unlinked) however we still observed the same correlations shown in Figure 4.3.

As previously discussed, housekeeping genes have been shown to be under stronger levels of purifying selection than other classes of genes [177], and it has been hypothesised that this is a result of their broad expression, intracellular location and key role in cellular processes. As shown in Figure 4.2 open chromatin is enriched with broadly expressed genes. We would expect this enrichment of housekeeping genes in relatively open regions of the genome, as open chromatin is likely to provide a more conducive environment for transcription. However the lower average dN/dS observed in open chromatin may simply be a consequence of this higher number of housekeeping genes in these regions. The exclusion of housekeeping genes from the analysis however has little affect on the correlations in Figure 4.3. Even the exclusion of all genes from the analysis whose 5' end is associated with a CpG island (which includes almost all housekeeping genes [178]) does not lead to the loss of the correlations between chromatin structure and dN, dS and dN/dS. In fact the rate of dN in CpG island genes, unlike that in genes not associated with a CpG island, is relatively constant across chromatin categories and does not show a significant correlation with chromatin. Consequently selection appears to maintain similar levels of dN in genes associated with a CpG island irrespective of their local chromatin structure.

To ensure these results were not confounded by CpG associated or sex chromosome specific factors (sex chromosomes have been shown to display abnormal rates of divergence when compared to the autosomes [161]), we calculated divergence rates at non-CpG, fourfold degenerate sites in genes on autosomes only. We also used human-chimp alignments instead of human-mouse alignments as the chromatin structure of the chimp genome should be more similar to that in humans, due to the dramatically closer evolutionary distance between human and chimp than human and mouse, and consequently the species of origin for each change should be less important (it should be noted however that even comparisons of human and mouse chromatin structures have shown striking levels of conservation even when the underlying DNA sequence has diverged [179]). However, as shown in Figure 4.4, the highest rates of divergence are still observed in genes from the most open regions of the genome.

4.3.3 Genes in closed chromatin display the highest levels of selection at synonymous sites

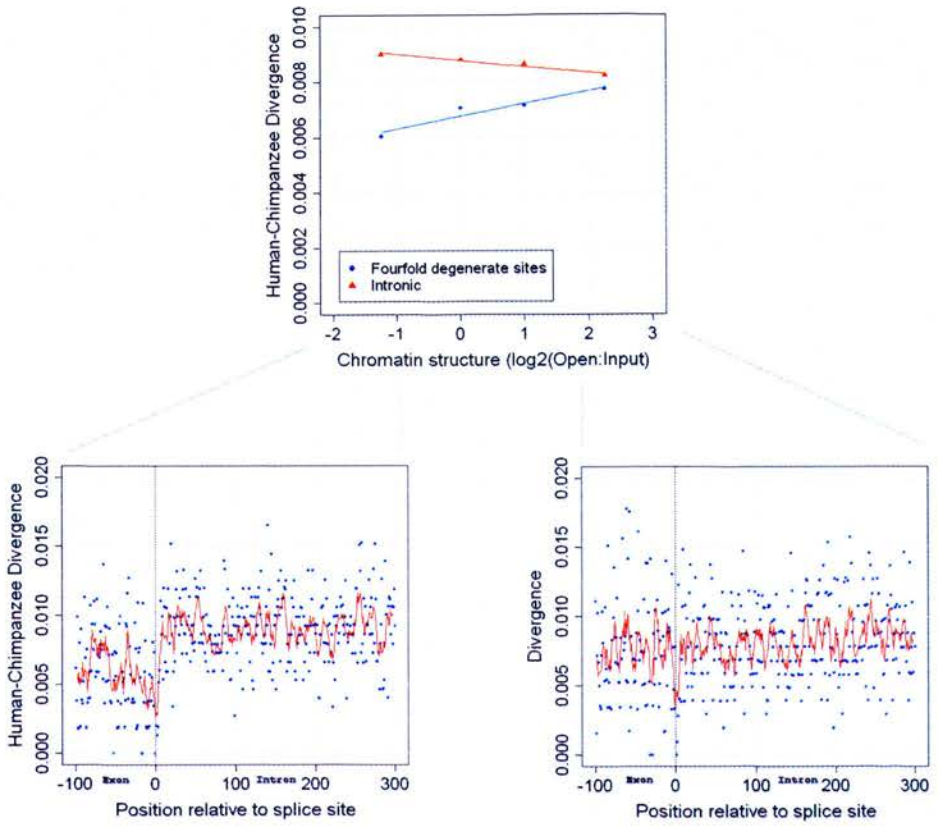


Figure 4.4: Intronic and exonic human-chimp divergence across chromatin categories. The divergence at non-CpG fourfold degenerate and intronic sites on autosomes only, with the divergence observed across the splice sites of the most open and most closed genes shown below. (closed exonic vs closed intronic Mann-Whitney U test: p = 4.4e-16; open exonic vs open intronic Mann-Whitney U-test: p = 0.053)

Although, historically the synonymous substitution rate (dS or Ks) has been used as a measure of the rate of mutation, there is increasing evidence that at least weak selection may be occurring at synonymous sites [175]. For example, it was shown in the chimpanzee genome paper that the rate of human-chimp divergence in exons was 25% lower than at neighbouring introns, and it is hypothesised in this paper that this is a result of direct purifying selection at synonymous sites [161]. To investigate the potential role of any selection on synonymous sites in the disparity

between dS and other measures of mutation, we analysed the rates of divergence observed across intron-exon boundaries, as in [161]. As shown in Figure 4.4 the rates of intronic divergence in open regions of the genome are comparable to those observed at corresponding exonic, fourfold degenerate sites. This would suggest that genes in open chromatin display little if any evidence for selection at their synonymous bases. However genes in closed chromatin display markedly higher rates of divergence at their intronic sites than at corresponding fourfold degenerate sites. Genes in closed chromatin therefore, unlike those in open, display strong evidence for synonymous site selection.

Although the rate of selection against both synonymous transitions and transversions is highest in closed chromatin, only the rate of synonymous transversions is strongly positively correlated with chromatin structure. The rate of transversions at fourfold degenerate sites shows no obvious trend across chromatin categories and consequently selection against transversions, unlike transitions, appears to be independent of any factors associated with chromatin structure (Figure 4.6). We are not aware of any reason for the observed association between rates of transitions at non-CpG fourfold degenerate sites and chromatin structure, but it could reflect constraint in motifs whose distribution is not uniform across the genome.

As previously shown open regions of the genome are particularly gene dense whereas closed regions are relatively gene poor [113]. Consequently, the use of dS as a measure of mutation rate may be appropriate for a large proportion of genes. However the use of dS as a surrogate measure of mutation rate for genes in closed chromatin will lead to the under-estimation of the true mutation rate in these regions and also the miscalculation of the levels of selection when used to measure dN/dS. Although dN/dS is a general indication of selective constraint measured by comparing substitutions between species other measures of selection are likely to be affected in a similar fashion. For example an alternative test for selection is to compare the synonymous and nonsynonymous differences between species with the synonymous and nonsynonymous changes observed within a species [180]. Measuring selection using data generated by Bustamante et al. [159] we compared the numbers of genes in each chromatin category displaying evidence for purifying selection. As can be seen in Figure 4.5 the proportion of genes with high P- (the posterior probability that a gene harbours excess amino acid variation) is substantially higher in open chromatin than it is in closed. However as the McDonald-Kreitman test uses synonymous sites as a measure of neutral mutation in the same way as dN/dS these results are likely

to also result from the selection at synonymous sites observed in this analysis.

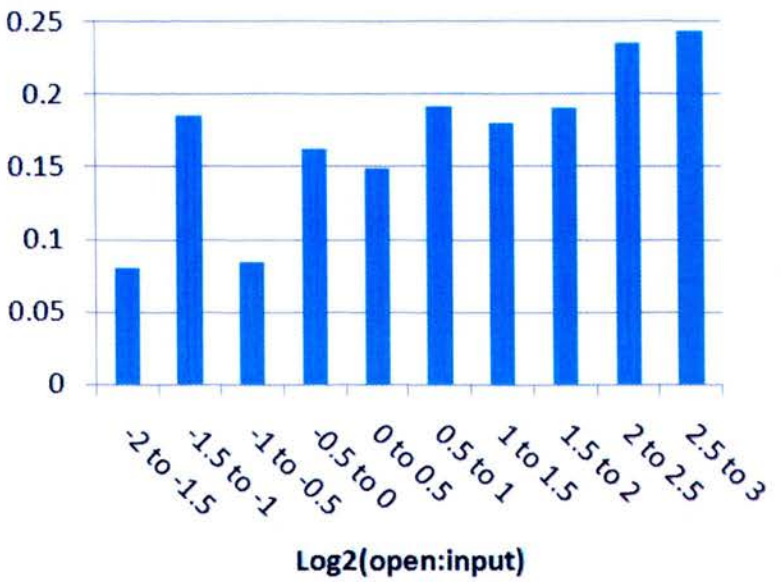


Figure 4.5: The proportion of genes across chromatin categories displaying strong evidence of purifying selection (i.e. whose P- was greater than 0.95)

4.3.4 Exonic Splice Enhancers and RNA secondary structure

It has been proposed that synonymous sites may experience constraint because they play a role in controlling splicing or RNA stability [175]. For example synonymous sites may be part of an exonic splice enhancer (ESE) motif or lead to a more stable base-paired RNA that is less susceptible to degradation. Although codon usage bias resulting from unequal abundances of tRNAs and subsequent selection at synonymous sites in favour of codons corresponding to the most abundant tRNAs has also been proposed as an explanation of synonymous site selection, the evidence for this in mammals is weak [181]. We therefore looked at the distribution of each predicted ESE motif across chromatin categories to see if their relative densities could explain the high levels of synonymous selection in closed chromatin. Although the density of a large proportion of ESE hexamers (44%) did display a significant negative correlation with chromatin structure (i.e. $p < 0.05$); given the base composition of ESE hexamers and coding regions across chromatin categories, we actually observed far fewer hexamers displaying a negative correlation than we would expect by chance

(66%). This is because coding sequence base composition is itself correlated with chromatin structure and ESEs also show biases in their base composition. As shown in Figure 4.6 excluding all sites from coding regions that overlap a predicted ESE hexamer leads to only a small increase in the rate of transitions observed at fourfold degenerate sites. Consequently, either there are many ESE motifs that are yet to be determined, or selection at synonymous sites is at most only partly the result of exonic splice enhancers.

We also looked at the distribution of gene types across chromatin categories; if genes whose RNA structure are important were preferentially located in closed chromatin we may expect an over-representation of non-protein coding genes in closed regions. As shown in figure 4.6 certain classes of non-protein coding genes are indeed over-represented in closed chromatin (rRNAs and snRNAs), while the distribution of other types of genes such as miRNAs and snoRNAs show no relationship with chromatin structure.

Further analysis is therefore required to determine why protein coding genes in closed regions of the genome display such comparatively high levels of selection at their synonymous sites. If it is indeed because of a requirement for a more stable secondary structure, then we may expect that the predicted stability of mRNAs from closed regions would be greater than those in open [182]. Future tests of this kind may help determine the reasons behind the enrichment of selection at synonymous sites in closed chromatin observed in this study.

4.3.5 Levels of linkage disequilibrium are also correlated with chromatin structure

It has been shown that mutation and recombination rates covary in the human genome [173]. We therefore used HapMap [183] genotype data to investigate whether fewer SNPs in regions of a more closed chromatin structure displayed evidence for linkage disequilibrium. By using Gabriel et al.'s [169] definitions of which SNPs show "strong evidence for historical recombination" (one sided upper 95% confidence bound of D' less than 0.9) and which are in "strong LD" (upper 95% confidence bound of D' greater than 0.98 and lower bound greater than 0.7) we determined that the proportion of pairwise comparisons displaying strong LD was consistently lower in closed chromatin (Figure 4.7). Likewise the proportion of pairs displaying strong evidence for recombination was highest in open chromatin. This was observed across

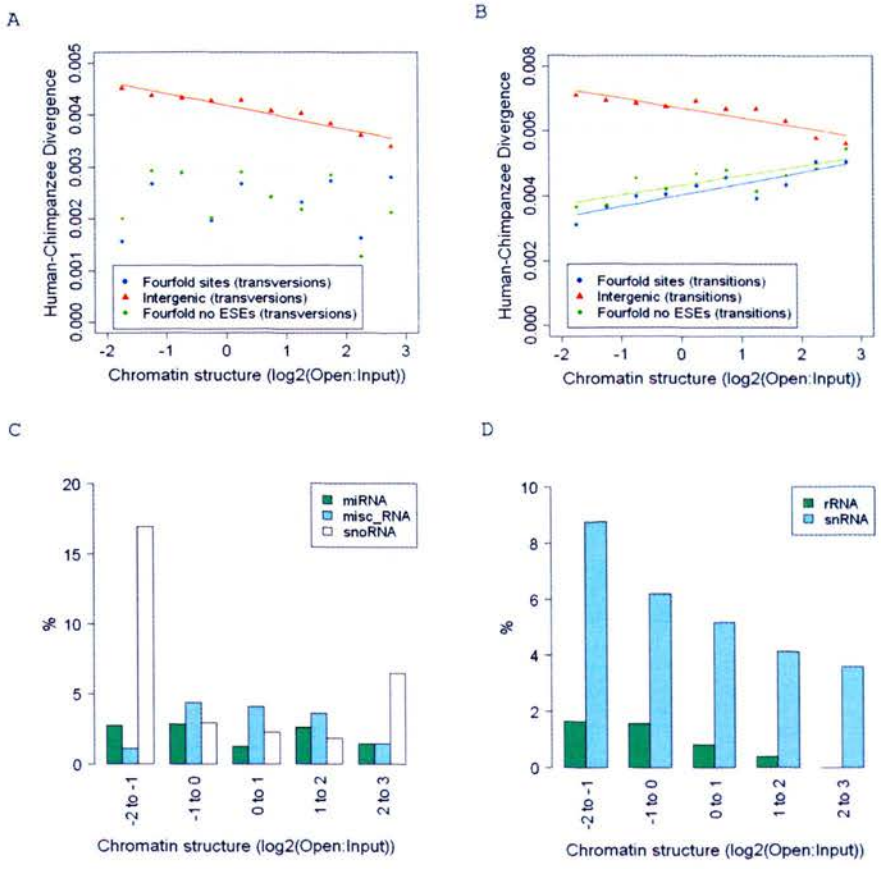


Figure 4.6: The effect of ESEs on fourfold degenerate site divergence and the ncRNA gene distributions observed across chromatin categories.
(A+B) The observed rate of transversions and transitions respectively, at fourfold degenerate site with and without ESE sites excluded (Fourfold degenerate site transversions: r^2= 0.02, p = 0.69; fourfold degenerate site transversions at non-ESE sites: r^2= 0.13, p = 0.30; intergenic transversions: r^2= 0.92, p = $1x10^{-5}$. Fourfold degenerate site transitions: r^2= 0.78, p = 0.001; fourfold degenerate site transitions at non-ESE sites: r^2= 0.67, p = 0.004; intergenic transitions: r^2= 0.81, p = $4x10^{-4}$). (C+D) The percentage of genes in each chromatin category that are of each Ensembl ncRNA class. Only the distributions of rRNAs and snRNAs show a significant negative correlation with chromatin structure (rRNA: r^2=0.96, p=0.004; snRNA: r^2=0.92, p=0.01)

all four HapMap populations. Recombination rate is not however the only biological determinant of LD. Rates of mutation and gene conversion also contribute to its disruption [169]. To investigate the potential contribution of recombination on this correlation we assumed that, if many SNP pairs are examined, the contribution of mutation rate and gene conversion should be approximately the same irrespective of the distance between markers (this may not be strictly the case when the markers are very close together but should hold across most distances). Therefore, a higher rate of LD disruption with increasing distance between SNPs should primarily be the result of increased levels of recombination (the greater the distance between markers the greater the chance of a recombination event). By comparing clones with a relatively closed chromatin structure to all clones in the 1Mb clone set we showed that the rate of increase of LD disruption was significantly lower in these closed regions than in the genome in general. This was observed in all four HapMap populations. Open chromatin did not however show a significant difference from the genome in general, suggesting that a closed chromatin structure may inhibit recombination but an open structure does not promote it. We therefore used the more statistically rigorous technique of McVean et al. [184] to estimate recombination rates and confirmed that recombination rates are significantly lower in closed chromatin than in the genome in general (Mann-Whitney test rate in closed clones versus all: $p = 0.031$) Extrapolation of the lines in Figure 4.7 to a distance of 0 between markers (i.e. no recombination) suggest that at most 5% of SNP pairs show "strong evidence of historical recombination" for reasons other than recombination (recurrent mutation, gene conversion, genotyping errors etc). There is also no significant difference between closed chromatin and the genome in general at this point, suggesting the observed correlation between chromatin and LD is not the result of the correlation between mutation rate and chromatin.

4.4 Conclusions

We have shown that rates of mutation and synonymous selection are correlated with chromatin structure. Regions of open chromatin display the lowest mutation rates and the least constraint at the synonymous sites of genes. Consequently previous observations of mutational hotspots in the human genome, high mutation rates around classes of genes involved in extracellular communication, the low dN/dS observed in

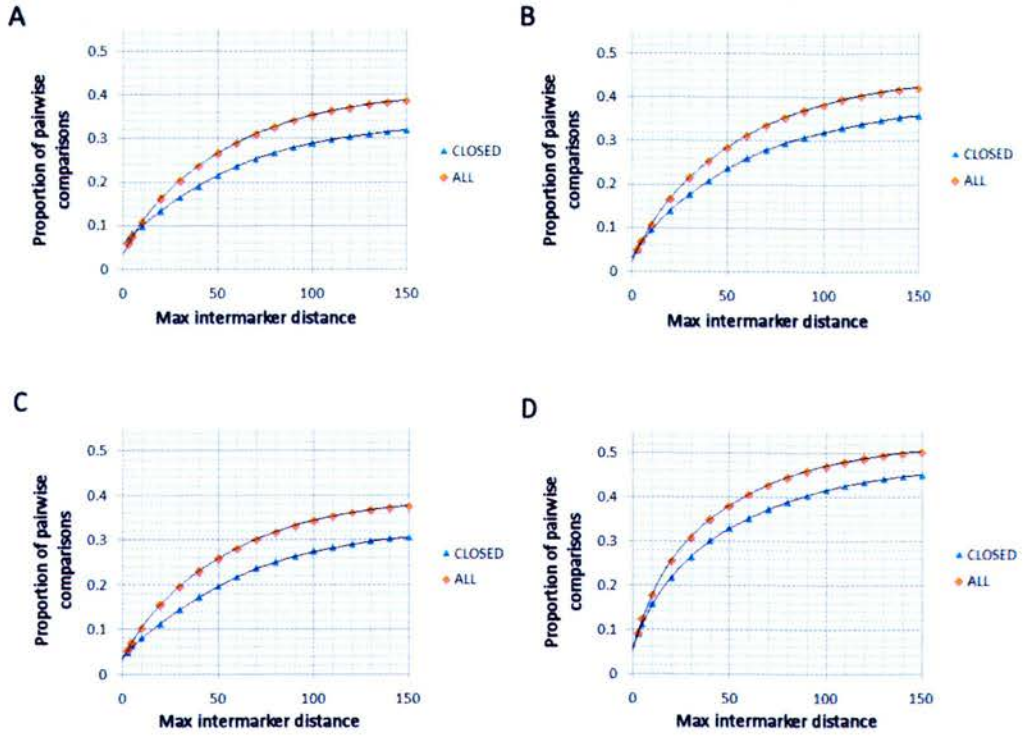


Figure 4.7: The proportion of SNPs, at various distances apart, that show strong evidence for recombination in the four HapMap populations. (A) CHB, (B) CEU, (C) JPT, (D) YRI.

housekeeping genes and the clustering of genes with similar divergence levels can all also be associated with chromatin structure. These correlations are observed despite the relatively low resolution of the chromatin dataset. The average length of the clones used in this analysis was 146kb, however the average human exon for example is approximately a thousand times smaller. There is consequently a disparity between the DNA regions whose rate of change we are measuring and the region of which we know the chromatin structure. The ability to measure chromatin structure at a higher resolution in the future may help increase the strength of these observed correlations.

We believe the lower mutation rate observed in open regions of the genome in this study is likely to be a result of these regions being more accessible to repair mechanisms, and not a result of selection. Indeed it is known that sites of transcription-coupled repair are clustered in the gene dense (and therefore) open chromatin regions of the genome [185]. Despite it now appearing that the majority of the genome is transcribed, Surralles et al. showed that transcription-coupled repair was preferentially located at these GC rich and gene dense regions (probably as a result of the higher levels of transcription in these areas). Likewise chromatin remodeling is known to be a precursor to DNA repair, and efficient DNA lesion detection is associated with relaxed chromatin structures [186, 187, 188]. However, contrary to mutation rate, we believe it unlikely that chromatin structure mediates selection on synonymous sites directly. Rather, it is more likely that genes that display a high level of selection at their synonymous sites are preferentially located in closed regions of the genome. It may be that these genes in general require especially tight transcriptional regulation, with a consequence being they are less accessible for DNA repair.

Chromatin structure is likely, however, to be only one of a number of factors that are associated with the variance in divergence rates observed across the human genome. This is supported by the fact that the levels of intergenic divergence of chromosome 19 are substantially higher than other autosomes, despite being gene dense and relatively open in structure. Most notably, both the chromatin dataset used in this analysis, as well as nucleosome formation potential [189], have previously been associated with GC content. Although this agreement between the lymphoblastoid chromatin dataset used in this analysis and other more general datasets is reassuring, GC content has previously been associated with rates of mutation and selection. However, although the mechanisms underlying the appearance of GC variability and

isochores along the human genome remain controversial, it has been proposed that they may be a result of selection for the structural requirements of DNA. For example, an increase in GC content has been associated with an increase in bendability of DNA and a decrease in curvature, properties associated with more open chromatin [189]. Further analysis is consequently required to determine the complex interplay between the various factors involved in rates of mutation and selection across the human genome.

These results have important implications for cancer. Approximately 75% of colorectal cancer cases are sporadic in nature, i.e. there is no indication that the disorder was inherited, instead it likely arose as a result of novel mutations and/or environmental factors [190]. Our analysis indicates that genes located in closed regions of the genome are particularly susceptible to such uncorrected mutations (although the rate of mutations was determined from fixed germline changes various lines of evidence, not least the incredibly strong relationship between chromatin structure and GC content that is fixed across tissues [113], suggest higher order chromatin structure does not change dramatically between cell types). Consequently not all cancer associated genes are equally likely to accumulate a novel mutation over equal periods of time. One of the most important genes in colorectal cancer, the "gatekeeper" APC, has been shown to be mutated in approximately 80% of sporadic colorectal tumours [4]. This gene is also located in a particularly closed region of the human genome. The local chromatin structure of APC may therefore be indirectly affecting the incidence of colorectal cancer. Work carried out in our lab involving the sequencing of APC in cell lines, has already shown that the background mutation rate of APC is higher than that at other randomly chosen genes, tentatively supporting this hypothesis [191].

4.4.1 High resolution chromatin dataset

The microarray used in the analysis above was the Sanger 1Mb array. However towards the end of my PhD the Bickmore lab produced a higher resolution 30k chromatin dataset and we were keen to see how results compared with this new data. After filtering out overlapping clones, the 30k dataset contained 4,988 remaining clones. However, although this dataset contained twice as many clones as the original 1Mb dataset, the correlations observed above were generally weaker with this new data. In all comparisons the largest discrepancies between the two datasets were at

the most open clones.

There were therefore two potential explanations; the first was that the new dataset was simply more accurate and the strong correlations we observed with the 1Mb dataset were an artifact of the relatively low resolution. However stronger correlations as a result of lower resolution seemed unlikely. Alternatively there may have been some error present in one of these datasets.

In order to further investigate these differences we compared the overlap between the 30k and 1Mb datasets. Although there is little direct overlap between the clones used in these studies (only 1 filtered clone was represented on both arrays), 858 clones of the new dataset showed greater than 99% overlap in its genomic position with a clone represented on the 1Mb array (by taking only those clones with a greater than 99% overlap it is unlikely that any differences observed between the platforms would be a result of differing G+C content of the clones). We therefore compared the $\log_2$(open:input) values of these clones, and as would be expected, there is a positive correlation between the platforms ($r^2 = 0.09$, $p < 2.2\text{x}10^{-16}$). The Bickmore lab had however also produced $\log_2$(closed:input) values for the 30k array, i.e. they had looked for enrichment for closed, as well as open, chromatin across the genome. We expected that in general there would be an inverse correlation between a clones $\log_2$(open:input) and $\log_2$(closed:input) values, and comparisons between this closed analysis and the 1Mb open dataset did indeed show such a negative correlation ($r^2 = 0.03$, $p = 2.0\text{x}10^{-7}$). However comparisons between the new 30k open and closed values showed a weak positive correlation ($r^2 = 0.012$, $p = 0.0012$). This is contrary to what we would expect and suggests that the clones most enriched with open chromatin are also the most enriched with closed chromatin. Although this is possible we believe the weaker correlations observed between mutation, selection etc and chromatin structure with the new 30k open dataset is a result of some incorrectly annotated clones, and mainly those clones are annotated as particularly open. We can see some evidence supporting this hypothesis in fig 4.8. Clones with a $\log_2$(open:input) value greater than 2 in the new dataset show a range of values in the 1Mb analysis, including as low as -1.5. However, conversely, the few clones with a $\log_2$(open:input) value greater than 2 in the 1Mb analysis all have positive values in the new dataset.

To determine those clones which showed agreement across platforms we determined the distribution of clone $\log_2$ values both within and across platforms. The distributions seen within each platform showed a high level of agreement with one

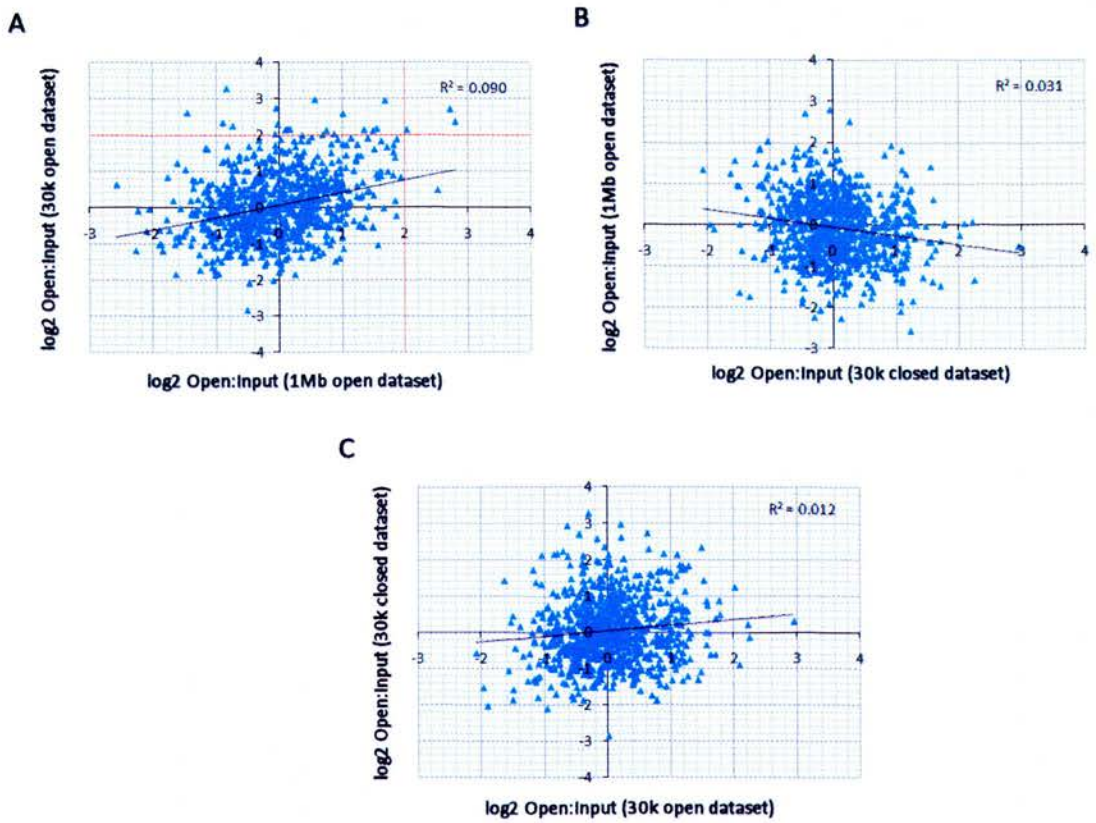


Figure 4.8: Chromatin dataset comparisons.
(A) Comparison of open datasets. (B) Comparison of old 1Mb open dataset to new 30k closed dataset. (C) Comparison of new open and closed datasets.

another and consequently we were able to assume all values, from all platforms were from the same distribution. We were therefore able to calculate the probability of obtaining two values as similar as observed for each clone. To compare the closed platform to the open datasets we reversed the closed distribution under the assumption that generally clones enriched with open chromatin will be depleted with closed chromatin. Although we observed differences in the distribution and clustering of clones showing strong agreements between platforms, for example almost a quarter of clones on chromosome 2 and 16 show strong agreement ($p<0.05$) between the open platforms compared to only 3% of clones on chromosomes 7 and 17, no biologically plausible explanation for differences between the platforms could be found.

Chapter 5

Candidate Gene Association Study

5.1 Introduction

5.1.1 Association Studies

Although various genes and pathways that participate in the development and progression of colorectal tumours have been identified, as mentioned previously there still appears to be a large number of genetic factors involved in colorectal carcinogenesis that have yet to be determined. There is therefore a requirement to try and identify further genes that may be involved in the development of colorectal cancer.

To date, the successful identification of genes involved in a particular disease has largely relied upon traditional positional cloning techniques, for example the use of linkage studies to look for co-inheritance of chromosomal regions with disease [192]. However, despite the success of these techniques with rare, monogenic conditions the complex genetic nature of cancers means this approach is unlikely to be sufficient.

To fully investigate the genetic influence of colorectal cancer would require identifying all differences observed between the genomes of affected individuals and those of controls. In this way polymorphisms, epigenetic states, gene copy numbers etc that contribute to the disease would be expected to be observed at different frequencies in the two groups of individuals; the basis of association studies. Although such a comparison on a sufficiently large number of individuals is currently unfeasible, much of the difference between individuals can be identified by determining only a subset of the variation among people. This is due to the fact that when genetic factors such as polymorphisms or gene copy number changes first arise in a popula-

tion they do so on a specific background. For example, when a single base mutation arises in an individual, the single nucleotide polymorphisms around it will each be of a certain allele. As this mutation is passed from one generation to the next it will be inherited with the same alleles around it, and consequently if an individual contains this mutation you will know the alleles of each of the surrounding polymorphisms (without having to type them). Over time the mutation may spread through the population to become detectable as a SNP itself and recombination and mutation will change the alleles around it and break up this relationship between the mutation and polymorphisms surrounding it. However, recombination between the mutation and the polymorphisms close to it will be rare so that the mutation will most often be seen with the same specific alleles of the surrounding SNPs. The original mutation and these SNPs are therefore in linkage disequilibrium and the combination of alleles observed with the mutation a haplotype. Linkage disequilibrium is therefore a strong tool in disease mapping and association studies as it allows only a subset of markers to be typed to capture the majority of the variation between individuals in a region.

So what is the advantage of association studies over traditional linkage analysis The answer is power. Detection of an allele with a relatively low relative risk using linkage analysis requires significantly more individuals to be tested than through the use of an association study. For example, computer models predict that to detect a disease allele with a genotypic risk ratio of 2 and a frequency of 0.1 would require 5382 families. Detecting this allele through an association study of affected siblings would require only 264 sib pairs [193].

Although the most powerful association studies are performed on families (i.e. with sib pairs or affected individuals and their parents), there are a number of advantages of using population-based cases and controls in association studies. Firstly, the role of a gene in cancer is often only apparent upon exposure to certain environmental factors. For example variations in the activity of glutathione s-transferases, enzymes involved in the detoxification of environmental carcinogens, are believed to affect susceptibility to tumours with large environmental contributions [194]. Family-based studies however are generally poor at assessing gene-environment interactions. Likewise, as analyses of family based studies such as TDT (the transmission/disequilibrium test) are based on comparing observed allele frequencies to those expected from parental genotypes, they can often be confounded by alleles which affect fertilisation success [195]. Association studies based on cases and con-

trols drawn from the population are also generally easier, quicker and cheaper to construct, especially in late onset diseases such as cancer where relatives of affected individuals are often deceased. On the other hand population based case/control studies are strongly affected by ethnic admixture [193]. For example subpopulations that differ in disease prevalence and allele frequencies can lead to artefactual associations. Properly matched cases and controls are therefore vital in non-family based association studies.

There has already been a number of notable successes in using association studies to map disease genes/variants. For example in 2004, Begovich et al. typed 87 putative functional SNPs in 475 individuals with rheumatoid arthritis and 475 matched controls [196]. A variant in *PTPN22*, a gene responsible for negative regulation of T-cell activation, was shown to have an odds ratio of 1.65 (95% confidence interval 1.23-2.20). This association was subsequently confirmed in a replication study of 463 affected individuals. Recently there has even been successes in relatively complex diseases. Three independent studies investigating age-related macular degeneration identified the same gene as associated with the disease [197, 198, 199]. Similar success in cancer has however never been certain.

The primary concern of using association studies to determine genetic factors influencing cancer is that unknown common cancer susceptibility genes that confer substantial risk may not exist. The ultimate aim of any cancer association study is to identify genetic variants that will lead to a clinical benefit. Rare variants that only affect a small subset of the population, though biologically informative, will never be feasibly tested in a population screen. For screening to be viable the variant must be relatively common and confer a reasonable relative risk. However recent studies suggest that such variants may not even exist in cancer. The environment is believed to make a substantial contribution to colorectal cancer risk with Lichtenstein et al. [42] estimating a 70% contribution of non-heritable factors. Likewise migration studies show that the incidence of cancer can change dramatically within one or two generations; too quick to be the result of the introduction of new susceptibility genes [200]. This combined with the numbers required to confidently detect variants of a realistic relative risk of 1.25 (>2000 cases) suggests that the search for cancer susceptibility variants is far from guaranteed to succeed.

Without testing cancer however it is impossible to know whether common cancer susceptibility variants of reasonable risk do or do not exist. The aim of any LD-based association study should simply be to be rigorous enough to detect them

if they do. Although various genetic markers can be used in LD-based methods of disease mapping, single nucleotide polymorphisms are relatively common in the genome (there are currently around 5 million validated Ref-SNP clusters in dbSNP) and therefore, despite generally only having two or three alleles at each SNP, have become the predominant marker used in association studies.

However, to take advantage of linkage disequilibrium in designing association studies it is first necessary to know which SNPs are in LD. To this end the SNP Consortium [167] and subsequently the International HapMap project [170, 183] were launched to not only discover SNPs but to also determine their genotypes in people from four different populations. The four populations chosen by the HapMap consortium were 45 Han Chinese from Beijing, 44 Japanese individuals from Tokyo, 90 individual (30 parent-offspring trios) from Ibadan, Nigeria and 90 individuals (again parent-offspring trios) from Utah. From these genotypes researchers are able to determine which SNPs are in LD and which capture most of the variation in a genomic region. The northern European descent of the individuals from Utah means the LD structure in this group is most likely to closely represent that of the Scottish population.

In 2005 HapMap released their phase I data with one SNP genotyped approximately every 5kb [183]. Examination of 10 500kb fully sequenced regions scattered across the genome illustrated that 80% of common SNPs are in perfect LD with another polymorphism (i.e. are completely redundant). It is not necessary however for tagging SNPs in association studies to be in perfect LD with the SNPs they are to capture (given enough cases and controls) and greater than 90% of variants were correlated with an r^2 of at least 0.8 with another SNP. Likewise comparison of these ENCODE regions to the phase I HapMap data suggests that approximately 75% of all common SNPs in the genome will be correlated with an r^2 of at least 0.8 with a polymorphism in the phase I release. Therefore picking HapMap tagging SNPs for an association study using an r^2 cutoff of 0.8 should capture 75% of all common variants in the genome.

The aim of this project was, as part of the SOCCS (Study Of Colorectal Cancer in Scotland)/COGS (Colorectal cancer Genetic Susceptibility Study) project, to prioritise, select and analyse SNPs that could be used to investigate any link between cancer and various candidate genes. These candidate genes included all known repair genes (as the majority of genes associated with colorectal cancer are involved in DNA repair) as well as several other genes that have been associated with tumours

of the colon (expression analysis, known interactor etc).

5.2 Methods

5.2.1 Gene selection

Repair genes were identified from two sources. The first, a list published by Wood et al. in 2005 [201], contained 145 manually annotated repair enzymes and genes associated with cellular response to DNA damage. This list, though subsequently updated, is available at `http://www.cgal.icnet.uk/DNA_Repair_Genes.html`. Each gene was mapped to an Ensembl gene id via its HGNC code [202]. The second source of repair genes was Gene Ontology terms [203]. The aim of the Gene Ontology project is to use controlled vocabularies to describe genes and gene products. GO terms are associated with genes by both manual curation and automatic annotation, for example by homology to known repair genes or domains. As the ontologies are arranged in a hierarchical fashion, the GO id for DNA repair as well as the id for all child nodes (such as mismatch repair) were determined. 107 Molecular function GO terms associated with DNA repair were identified and are listed in the appendix in chapter 9. All Ensembl genes associated with one of these GO terms (as of February 2006) or listed by Wood et al. are also shown in the appendix in chapter 9.

SNPs with a previous association with colorectal cancer were also identified from two sources. The first was a list published by Kemp et al. listing polymorphisms published as having been tested for an association with colorectal cancer [41]. SNPs annotated by Kemp et al. as having some or good evidence of an association with colorectal cancer were identified (i.e. all SNPs with reported but perhaps unconfirmed associations with colorectal cancer). The second source was a list of papers compiled by Dr Barnetson, each reporting an association between a variant and colorectal cancer. Insertions, deletions etc were ignored due to their difficulty of being typed using Illumina bead arrays. Each SNP, where possible, was mapped to a current dbSNP rs number. In total a list of 59 variants with a previous association to colon cancer was compiled (see appendix in chapter 9).

A further list of candidate genes was compiled containing genes with further associations with colorectal cancer. Each was mapped to an Ensembl gene id where

possible and these can be seen in the appendix in chapter 9.

5.2.2 SNP selection

All non-synonymous Single Nucleotide Polymorphisms located in our list of DNA repair genes were identified through dbSNP. Only SNPs annotated as validated through non-computational methods were selected. HapMap tagging SNPs were identified through the use of the Haploview program [57] that contains an implementation of Paul de Bakker's Tagger tag SNP selection algorithm [204]. As many SNPs can not be typed on the Illumina bead array many potential tagging SNPs had to be excluded. Likewise, as we were keen to compare our repair analysis to a similar study by Professor Gallinger at the University of Toronto, we preferentially selected 433 tagging SNPs typed in his study. We also forced non-synonymous HapMap SNPs to be used as tags where possible to minimise redundancy.

Only HapMap Phase I SNPs in our repair and candidate genes with a minor allele frequency of at least 5% were tagged in this study. The r^2 cutoff we used for selecting tagging SNPs was 0.8, i.e. every HapMap SNP within 10kb of one of our genes with a MAF greater than 0.05 had an r^2 of at least 0.8 with a tagging SNP. Due to the number of genes involved, the tag selection process was automated. This was achieved through a MySQL database and Perl scripting.

We compared the numbers of SNPs required to tag our genes using both pairwise and aggressive tagging (where aggressive tagging is where multi-marker tags are allowed). Although aggressive tagging required approximately 8% fewer tags, we did not feel this was sufficient to outweigh the more complex analysis required and the potential for missing disease haplotypes. Cases and controls were selected matched according to age, sex and postcode by Dr Tenesa.

5.3 Results and Discussion

5.3.1 Gene and SNP selection

The aim of this project was to design an association study, as part of the COGS/SOCCS project, with the aim of testing the association between colorectal cancer and various

candidate genes and SNPs. As discussed the main group of candidate genes being tested were the DNA repair genes, a list of which were compiled from two sources; GO term annotation and a paper published by Wood et al. in 2004. As can be seen in figure 5.1 the majority of repair genes were identified by both methods, with 116/220 genes being both associated with a repair GO term and listed by Wood et al. Far fewer genes were identified exclusively by Wood et al. only than solely through GO term annotations. However many genes are assigned a GO term through automated annotation and are therefore only predicted to be involved in DNA repair. This does mean however that these genes are less likely to have been previously tested for an association with cancer.

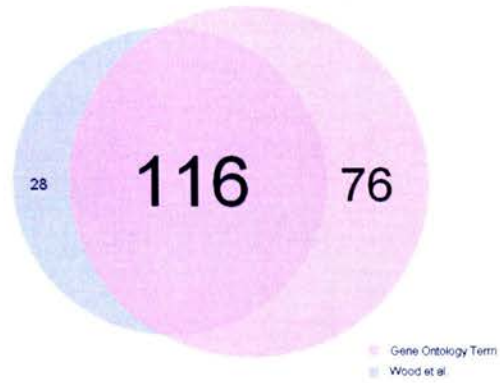


Figure 5.1: Overlap between repair genes listed by Wood et al. and those identified by GO term analysis

The second group of candidates contained various genes thought to be involved in cancer. These included the galectins shown to be differentially expressed in colorectal cancer (see chapter 1) as well as genes involved in folate metabolism and alcohol intake such as thymidylate synthase.

The tagging SNPs in both groups of genes were selected using Haploview. On average each gene contained 6 tag SNPs. However, as can be seen in figure 5.2 the distribution of tag counts per gene was skewed, a result of a few genes requiring many tag SNPs. The median tag count was 5. The average distance between tagging SNPs ranged from one tag every 691 base pairs in *GADD45G* to one tag every 81kb in *HMGB1*.

As can be seen in figure 5.3 the majority of the tag SNPs picked (54%) had no other proxies in the relevant repair gene and were therefore selected purely because

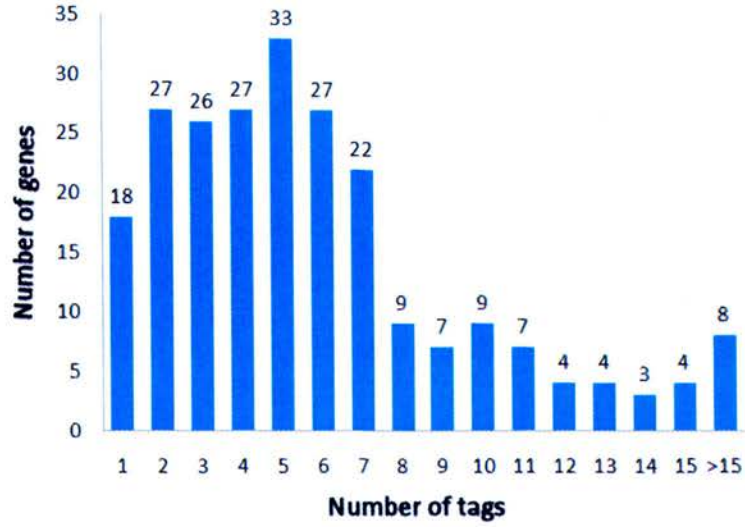
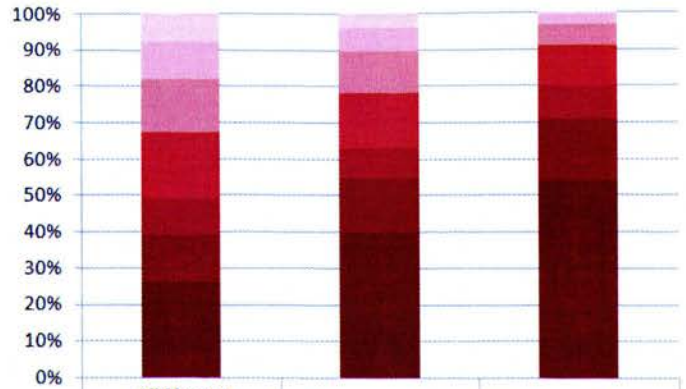


Figure 5.2: Number of tags per gene

they tagged themselves. However, although tag SNPs were selected because of their ability to tag the HapMap SNPs in repair genes, many of the tag SNPs picked also had proxies outside the immediate region of its gene. As shown in figure 5.3 only 40% of tag SNPs picked had no HapMap proxy within 500kb. This is compared to 26% of all HapMap SNPs in the genome having no proxy. However, this is as would be expected as tag SNPs will be enriched for SNPs that can only tag themselves (as all SNPs with no proxy have to be selected compared to, for example, only half the SNPs that have 1 proxy). The total number of SNPs tagged were ~2500 in repair genes, and ~6000 in the genome as a whole.

All validated, nonsynonymous SNPs located in one of our repair genes were also typed, along with a further 59 SNPs that had a previous association with colorectal cancer. Although this gave a total of 1920 polymorphisms, 149 SNPs failed genotyping and 148 were fixed in our population (table 5.1) (the majority of which were nonsynonymous SNPs from dbSNP that either did not in fact exist or were not present in this population). Two SNPs departed from Hardy-Weinberg equilibrium with a $p <= 0.001$, however this was no more than what would be expected given the number of SNPs tested in this study, and no SNPs departed with a $p <= 0.0001$. We were therefore left with 1623 successfully typed SNPs that showed variation within our population.



	All Phase I SNPs, whole genome	Our tags, whole genome	Our tags, just our genes
>20	7.67	4.01	0.80
11-20	10.65	6.49	2.41
6-10	14.37	11.30	5.82
3-5	18.34	15.18	11.24
2	9.80	8.36	9.03
1	13.22	14.98	16.92
0	25.94	39.67	53.78

Figure 5.3: Proxies per tag.
The percentage of HapMap SNPs that are within an r2 of 1 with 1, 2, 3-5, 6-10, 11-20 or greater than 20 other HapMap SNPs; the percentage of our chosen tag SNPs that are within an r2 of 1 with 1, 2, 3-5, 6-10, 11-20 or greater than 20 HapMap SNPs and the percentage of our chosen tag SNPs that are within an r2 of 1 with 1, 2, 3-5, 6-10, 11-20 or greater than 20 HapMap SNPs located within a repair gene.

Study	SNPs selected	Non-synonymous	Genotyped successfully	Fixed in population
Repair genes	1650	414 (25%)	1530 (93%)	138 (9%)
Candidate genes	211	26 (12%)	187 (89%)	7 (4%)
SNPs with prev. assoc.	59	30 (51%)	54 (92%)	3 (6%)

Table 5.1: SNP counts

As we used the CEU HapMap data to select our tagging SNPs we were keen to see how this population compared to our Scottish controls (for validation purposes). We therefore compared minor allele and heterozygote frequencies across populations. Figure 5.4 clearly shows that the frequencies observed in our Scottish controls is highly similar to that in the HapMap CEU population. Frequencies observed in the other HapMap populations are however markedly different. The only SNP whose allele and heterozygote frequencies were considerably different between our Scottish controls and the CEU data was rs2061783. This SNP had an observed minor allele frequency of 0.5 in the HapMap data but of only 0.023 in our Scottish controls. It is possible that this SNP does in fact show such differences between its frequencies and is to some extent a marker of the Scottish or Utah populations. However it seems more likely that it was simply mistyped in one of the two studies. Although it would have been possible to also compare LD structures across populations to ensure the validity of our tagging SNPs, previous work in our lab has shown that power in association studies is only marginally affected by the choice of HapMap population [205].

5.3.2 Comparison to Whole Genome Data

A number of the polymorphisms genotyped in this study were also typed as part of a whole genome association study carried out by our lab. As this study had been carried out on a second Illumina platform we were able to determine the reproducibility of our results. In total 1978 people and 765 SNPs had been typed in both studies[1]. As can be seen in figure 5.5 the reproducibility between people and SNPs was high, and of 1,055,434 genotypes, 99.95% were identical in both studies. The polymorphism rs2061783 that had shown such a marked difference in allele and heterozygote frequencies between the HapMap CEU and our Scottish population had a reproducibility of 100%, which suggests that the frequencies observed in our population for this polymorphism were real and not the result of incorrect genotyping. Some polymorphisms however were clearly mistyped in one of the studies. For example, rs1799949 had 236 individuals whose genotypes were different between studies and in all these cases the individual was typed as a GG in the whole genome study and as an AG on the custom array. Heterozygote and minor allele frequencies observed by

[1]More SNPs than we would have liked were typed in both studies as the Illumina 550k content was unknown when designing the custom array

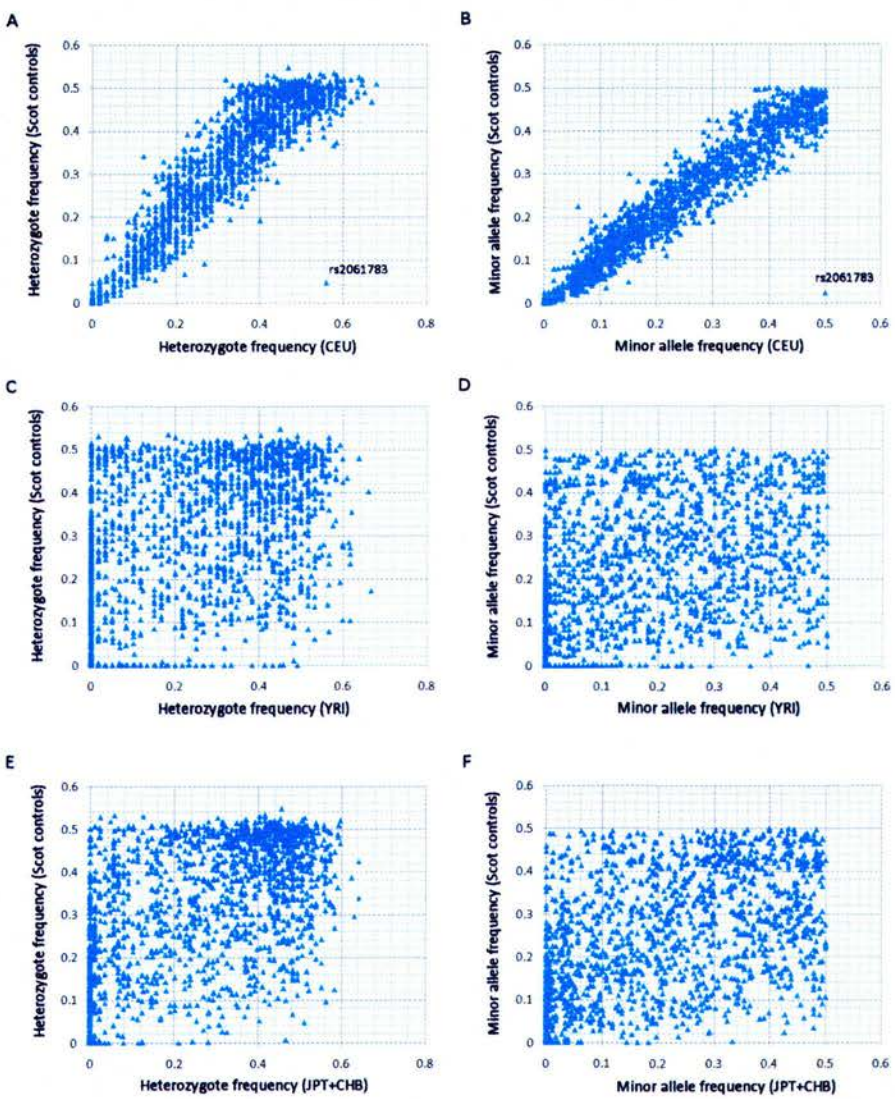


Figure 5.4: Control versus HapMap heterozygote and minor allele frequencies

Perlegen and HapMap in Europeans at this SNP are however more consistent with those observed in the whole genome than the custom study; allowing us to hypothesise this SNP was mistyped on the custom array. Some of the discrepancy between platforms was also likely the result of mislabeled individuals. For example, MD2246, showed a reproducibility of only 54%, it is however difficult to determine on which platform this individual was incorrectly genotyped. SNPs and individuals with low reproducibility percentages were removed from further analysis (People to exclude were determined by Dr Farrington, a threshold of 90% was used for excluding both SNPs and individuals).

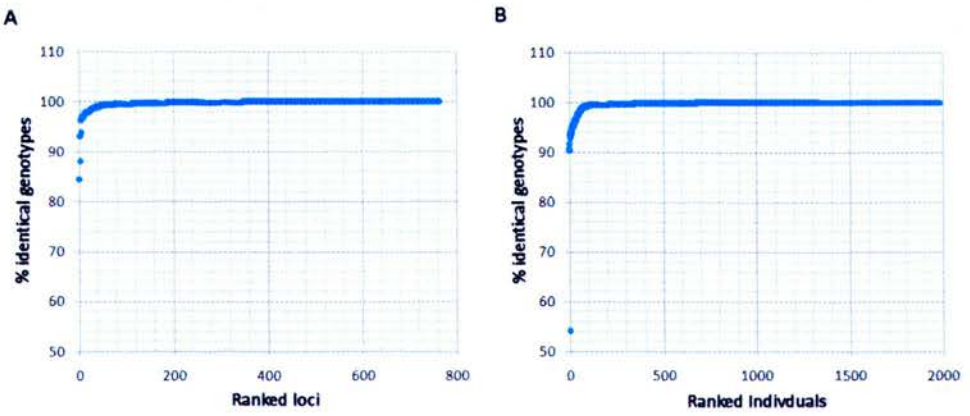


Figure 5.5: Genotyping reproducibility
Reproducibility between Illumina platforms for (A) the 765 SNPs and (B) the 1978 individuals successfully typed in both studies

5.3.3 Testing for association

To test for association between colorectal cancer and each SNP on the custom array a Chi-Squared test of allelic counts was used (i.e. a chi-squared test was performed on the counts of each allele at each SNP in cases versus controls, calculated using the PLINK program [206]). The top 20 polymorphisms with their corresponding p values are shown in table 5.2. Due to the number of tests performed it was necessary to correct these values for multiple testing array wide, and a number of strategies were used. These included the calculation of False Discovery Rates (q values, Benjamini & Hochberg and Benjamini & Yekutieli) and the application of Bonferroni, Holm and Sidak multiple test corrections. However in all cases, after multiple test correction,

Chr	dbSNP id	MAF cases	MAF cont	chi-sq	p value	OR	OR low95%	OR high95%	SNP type
10	rs10906777	0.04559	0.02595	11.1	0.0008615	1.793	1.266	2.54	Repair gene
18	rs8305	0.3181	0.2755	8.608	0.003347	1.227	1.07	1.407	Repair gene
18	rs483640	0.3177	0.2762	8.133	0.004347	1.22	1.064	1.399	Repair gene
4	rs373215	0.1793	0.1464	7.88	0.004999	1.274	1.076	1.51	Repair gene
3	rs3774332	0.06359	0.08683	7.658	0.005653	0.7142	0.5622	0.9073	Repair gene
6	rs9267803	0.2228	0.2595	7.246	0.007106	0.8183	0.7071	0.9471	Candidate gene
1	rs1010447	0.2541	0.2919	7.121	0.007617	0.8263	0.7182	0.9507	Repair gene
14	rs181564	0.3673	0.3283	6.623	0.01007	1.188	1.042	1.354	Candidate gene
14	rs11623756	0.3673	0.3283	6.623	0.01007	1.188	1.042	1.354	Repair gene
4	rs407555	0.2377	0.2041	6.496	0.01081	1.216	1.046	1.414	Repair gene
1	rs10864490	0.2215	0.256	6.449	0.0111	0.8271	0.7143	0.9577	Repair gene
8	rs6470522	0.1797	0.1499	6.387	0.0115	1.243	1.05	1.471	Repair gene
16	rs16260	0.2562	0.2919	6.344	0.01177	0.8353	0.7261	0.9609	Prev_assoc
1	rs11811771	0.4734	0.4336	6.298	0.01209	1.174	1.036	1.331	Repair gene
17	rs1799966	0.3064	0.3433	6.157	0.01309	0.8448	0.7394	0.9652	Repair gene
17	rs16941	0.3059	0.3427	6.086	0.01363	0.8453	0.7397	0.9661	Repair gene
2	rs7594702	0.2147	0.1836	5.97	0.01455	1.215	1.039	1.421	Repair gene
2	rs848292	0.3202	0.3568	5.914	0.01502	0.8491	0.7442	0.9688	Repair gene
1	rs3765767	0.5166	0.4775	5.905	0.01509	1.169	1.031	1.327	Repair gene

Table 5.2: Top twenty most significant SNPs

no SNPs approached significance. Under the null hypothesis p values should be distributed uniformly between 0 and 1 with any enrichment of low values due to true positives, and as can be seen from figure 5.6A there was no apparent enrichment of SNPs with low p values within our study.

As the unit of interest in this study was generally genes rather than SNPs, and therefore each gene rather than each SNP is really the unit of independence, we also corrected for gene wise type I error rates while also correcting for experiment wide testing by running 1000 (phenotype) permutations and only examining the top hit from each gene (we are only interested in whether any SNP in each gene shows an association with colorectal cancer and not all SNPs, for more details see PLINK set-based tests `http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml#set`). Although this technique is biased towards genes with relatively few SNPs, no genes were significant at greater than the 20% level.

Consequently having corrected for multiple testing no SNPs approached significance irrespective of the correction used. However one question we were keen to answer in this analysis was whether currently available SNP prioritisation programs such as SIFT [66] and SNPs3d [207] can be applied to cancer association studies to predict functionally important polymorphisms. It is thought that SNPs involved in cancer are less well conserved than their monogenic counterparts due to the complex gene-environment and late onset nature of the disease. Consequently, the power of programs that predict the impact of a polymorphism on disease through analysing conservation profiles is likely to be diminished, and SNP prioritisation programs may not therefore be applicable to cancer. However if polymorphisms associated with carcinogenesis are indeed generally more conserved than SNPs in general, then conservation may prove an alternative method for determining true positives from false positives.

To test this hypothesis we compared the uncorrected chi-square values from above to the output of various SNP prioritisation programs. Figure 5.7 shows the results of plotting the SNPs3d score of a non-synonymous SNP versus its corresponding chi-square value. As can be seen, there is no apparent relationship between these values, which suggests that Prioritisation of SNPs according to their SNPs3d score is unlikely to enrich for polymorphisms displaying an association with cancer.

However as shown in figure 5.7 those SNPs with high chi-square values generally showed low SIFT scores. Conversely, as expected, polymorphisms with low chi-square values showed a variety of SIFT scores (as many polymorphisms will be conserved

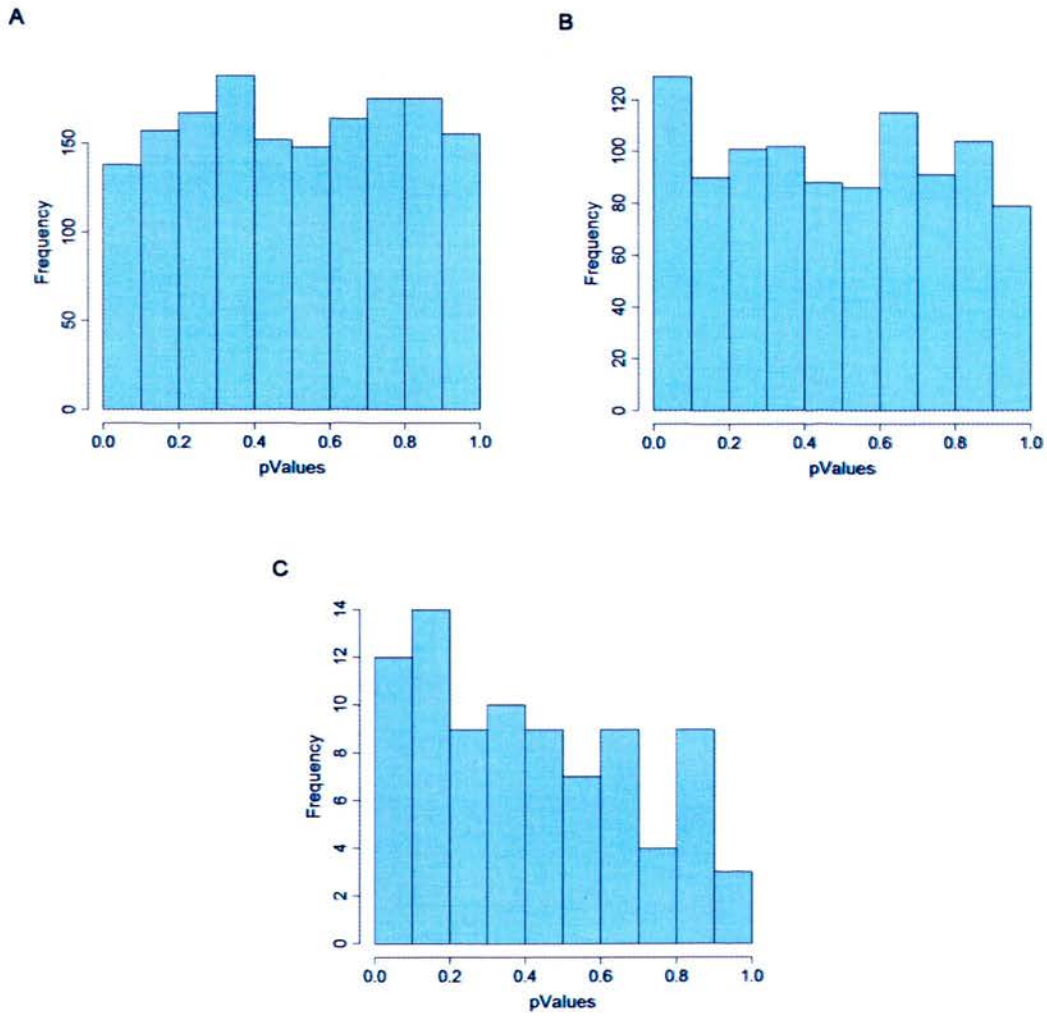


Figure 5.6: (A) The distribution of the p values of all the SNPs on the custom array. (B) The distribution of the p values of the Gallinger proxies of the custom array SNPs. (C) The distribution of the p values of the Gallinger proxies corresponding to the custom array SNPs with a p value less than 0.1

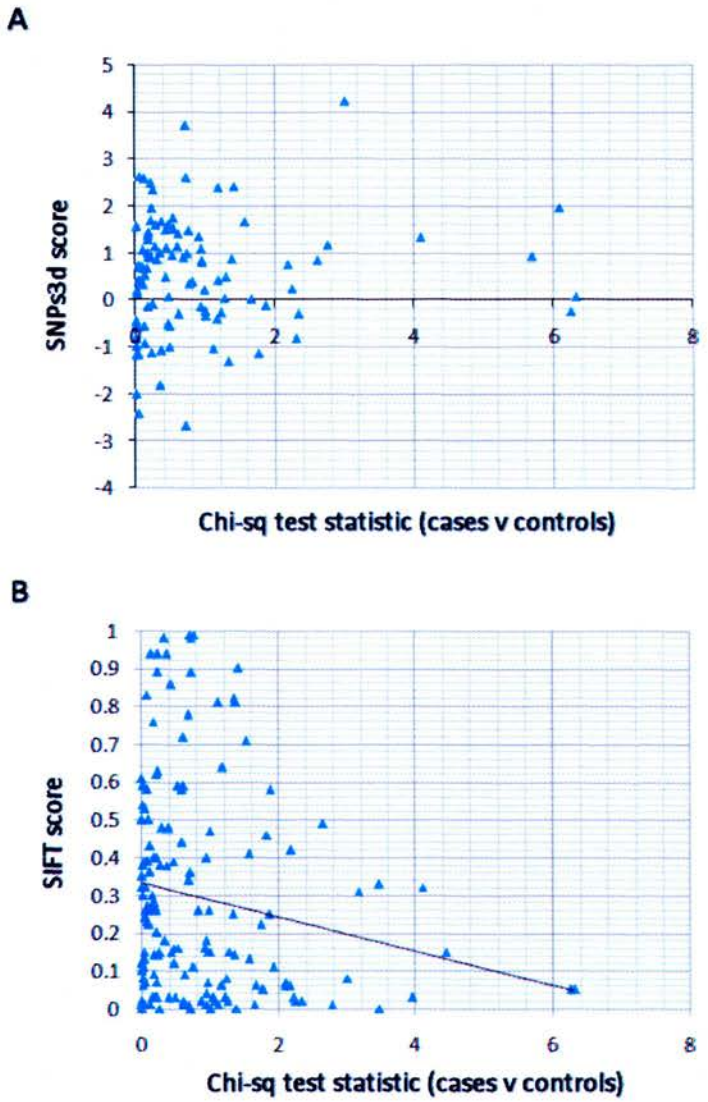


Figure 5.7: SIFT and SNPs3d scores against allelic case versus control chi-square values for our non-synonymous SNPs (SIFT versus chi-sq: r=0.18, p=0.026)

for reasons other than an involvement in cancer). This is in agreement with a more limited analysis by Zhu et al. that also observed a weak correlation between the odds ratios of 46 non-synonymous SNPs and their corresponding SIFT scores [208]. It is likely that the poor performance of SNPs3d relative to SIFT is at least partly a result of the fact that it is trained on monogenic diseases. SIFT on the other hand makes ab initio prediction with only the thresholds being determined from monogenic data. The threshold of 0.05, recommended by the authors of SIFT for annotating SNPs as deleterious, is likely to be too strict for use in cancer for the reasons discussed previously.

Despite this correlation between SIFT and chi square scores, the application of SIFT in designing association studies is likely to be limited. The modest enrichment in the proportion of detected true positives obtained is unlikely to outweigh the cost of missing further true positives considering the ever decreasing cost of SNP genotyping and the relatively small number of polymorphisms actually associated with cancer. However programs like SIFT may prove useful in separating true positives from false positives in the subsequent analysis of association studies. For example, we would expect true positives to in general be more conserved than false positives and for true positives to be more substantially enriched from within this group. The fact that there is an apparent relationship between conservation and chi-square value in our dataset suggests that there probably are indeed true positives within the set of non-synonymous polymorphisms tested, however their effect on colorectal cancer is limited and they can not be detected given the power of our study. (The polymorphism in figure 5.7 with the highest chi-square value, that is also highly conserved, is located in *BRCA1* a gene already known to play an important role in various cancers.)

Although our dataset is relatively sparse with SNPs of intermediate to high chi-square values, the fact that polymorphisms with intermediate chi-square values in figure 5.7 also show, at most, intermediate SIFT conservation scores, suggests that more SNPs may be associated with colorectal cancer than is at first apparent. It may be the case that some SNPs with relatively low chi-square scores show some level of conservation as they have only a minor role or they interact epistatically with other polymorphisms. We therefore tested the difference in SNP-SNP association between cases and controls both through PLINK and using logistic regression with an additive model (in both cases the results were similar). Although a number of SNP-SNP comparisons displayed small p values in this analysis (down to $7\text{x}10^{-7}$), given the

number of tests involved this is as would be expected. After Bonferroni, FDR and permutation multiple test corrections no pairwise comparisons were significant at greater than the 20% level.

5.3.4 Replication and comparison to further populations

If true positives do exist in a dataset they should replicate in other studies. As previously mentioned tags in our study had been preferentially selected that had previously been typed by the Gallinger group in Canada, and in total 433 SNPs had been typed by both groups in a total of of 2207 cases and 2211 controls. Likewise a number of our polymorphisms had been typed by the Houlston group in London in a whole genome scan of 1585 people (620 cases and 965 controls). Although this meant a reasonable proportion of SNPs could be directly validated in at least one further population, through the use of data from our own whole genome scan, we were also able to identify a further group of SNPs typed in the London population that were in high LD with one of our polymorphisms. In this way we identified 84% of our polymorphisms had been directly typed in the London dataset or could be assigned a proxy with an r^2 of at least 0.7. However substantially less SNPs could be validated using the Canadian data for two reasons. Firstly their genome coverage was more sparse than that of the Houlston group (they had only typed approximately 110k SNPs compared to over 500k typed in the London dataset). However perhaps more importantly we also had to rely on HapMap data to calculate LD between our SNPs and those typed by the Gallinger group. As 143 SNPs successfully genotyped in our study had not been typed in the HapMap CEU population, we were unable to identify tags for these polymorphisms. Consequently, to partly overcome this relatively sparse coverage of the Gallinger data we allowed both single and multi-marker tags to act as proxies (this was achieved by altering the Haploview source code so that the best single or multi-marker test was outputted for each SNP). In total 91% of our SNPs had been replicated (or pseudo-replicated by a proxy with an r^2>0.7) in at least one population and 53% were replicated in both.

If true positives existed within our dataset we would expect to see some enrichment of SNPs with low p values (the basis of an FDR analysis). However, as shown in figure 5.6A no such enrichment was observed. The proportion of true negatives estimated from this distribution using the qvalue R package was 99.5%. Our conservation analysis however suggested that this may be an over-estimation and that

some SNPs were indeed involved in cancer though their effects were probably small. Consequently to enrich for any true positives we identified those SNPs that had a p value of less than 0.1 in our analysis and that had been replicated in the Gallinger dataset. As can be seen from the distribution of Gallinger p values of these SNPs, an enrichment of low p values could now be observed (Figure 5.6C). Likewise the estimation of the proportion of true positives among this subset of SNPs was now at 53%, compared to 15% among the Gallinger dataset as a whole. Consequently the use of consecutive replication studies allowed us to enrich for potential true positives among our dataset.

As shown in tables 5.3 and 5.4, 55 SNPs (or their proxies) had a p value less than 0.1 in at least two populations examined, with three SNPs, rs12724233, rs11236164 and rs2247233, having a p value less than 0.1 in all three. Two of these SNPs, rs12724233 and rs11236164, are located in close proximity to known mismatch repair genes; the tumour associated *tp73* and the polymerase *POLD3* respectively. The third SNP rs2247233, like rs11236164, is associated with a DNA polymerase (*POLG*). It should be noted however that the best proxy of rs12724233 in the Gallinger population was in very limited linkage disequilibrium with the original SNP. This SNP is therefore a poor tag of rs12724233. However many SNPs in the region of tp73 displayed a low p value in the Gallinger dataset and we therefore included this region for further analysis. Figure 5.8 shows the odds ratios and corresponding confidence intervals of SNPs rs11236164 and rs2247233. If none of the SNPs in our dataset with an r^2 of at least 0.7 in our replication datasets were truly associated with colorectal cancer we would expect 1 SNP (or more precisely 0.85) to have a p value less than 0.1 in all three populations simply by chance. As we saw 2 it would be difficult to say there is much meaningful enrichment in this analysis.

Further SNPs also had a p value less than 0.1 in all populations where a suitable proxy was available. For example rs16260 was untaggable in the Houlston population and a had a relatively poor tag in the Gallinger dataset. However as it had previously been associated with colorectal cancer this SNP is likely to be a strong candidate for being associated with the disease.

Our SIFT analysis and examination of replication across studies consequently suggests that a number of our repair loci may indeed be involved in colorectal cancer. However our study is underpowered to detect them as a result of their attributable risk being small (OR <1.2). The growth in the number of association studies published has led to it becoming increasingly clear that the majority of disease variants

Chr	dbSNP id	Gene	SNP type	p		Houlston tag	r2	Houlston p		Gallinger tag	r2	Gallinger p
1	rs12724233	TP73/LOC339448	Repair	*0.01799*		rs12724233	1	*0.012*		rs4130091	0.023	*0.0508*
11	rs11236164	POLD3	Repair	*0.05848*		rs11236164	1	*0.0597*		rs11236164	1	*0.0138*
15	rs2247233	POLG	Repair	*0.09247*		rs2247233	1	*0.0195*		rs2238304,rs2351002	0.935	*0.0464*
16	rs16260	CDH1	Prev_assoc	*0.01177*		Untaggable	0			rs1562480	0.398	*0.064*
15	rs2072266	POLG	Repair	0.1078		rs2307449	0.988	*0.0063*		rs2351002	1	*0.0375*
1	rs3737589	TP73/KIAA0495	Repair	*0.05048*		rs3737589	1	0.101		rs10492942	0.077	*0.067*
10	rs4253077	ERCC6	Repair	0.1273		rs4253077	1	*0.079*		rs1917800	1	*0.0252*
6	rs9462085	PPARD	Candidate	0.1127		rs9462085	1	*0.0719*		rs10484578	0.579	*0.0472*
21	rs2835342	CHAF1B	Repair	*0.06031*		rs2835342	1	0.1256		rs2244034	0.291	*0.0631*
23	rs3088074	ATRX	Repair	*0.08846*		rs9781965	0.983	*0.0427*		Untaggable	0	
20	rs1047972	STK6	Prev_assoc	0.1661		rs1047972	1	*0.0311*		rs1047972	1	*0.0736*
17	rs2240308	AXIN2	Candidate	0.2168		rs2240308	1	*0.0583*		rs2240308	1	*0.0091*
12	rs812498	TDG	Repair	*0.08503*		rs812498	1	*0.1056*		rs1165670,rs2438112	0.371	*0.0936*
12	rs4883544	POLE	Repair	0.1944		rs4883627	0.996	*0.0185*		rs7966242,rs4883616,rs1132375	0.755	*0.088*
21	rs218631	CHAF1B	Repair	*0.02793*		rs218631	1	0.2101		rs2244034	0.132	*0.0631*
6	rs2395626	FANCE	Repair	0.2354		rs2395626	1	*0.0441*		rs10484578	0.354	*0.0472*
14	rs2064827	RAD51L1	Repair	*0.04307*		rs2064827	1	0.2653		rs8013026	0.066	*0.0507*
10	rs3750861	KLF6	Prev_assoc	*0.09803*		rs3750861	1	*0.0403*		rs7902434	0.075	0.2317
21	rs190068	CHAF1B	Repair	0.3102		rs190068	1	*0.0085*		rs2244034	0.27	*0.0631*
10	rs3793903	MGMT	Repair	0.2724		rs4751118	0.782	*0.0428*		rs3750825	1	*0.0688*
14	rs1713419	TEP1	Repair	*0.09076*		rs1713419	1	*0.0388*		rs7160770,rs938886	0.646	0.262
6	rs2267666	PPARD	Candidate	0.3447		rs2038068	0.955	*0.0178*		rs10484578	0.655	*0.0472*
2	rs768298	FANCL	Repair	*0.01603*		rs13411119	0.829	*0.0495*		rs768298	1	0.3579
17	rs2287321	RPA1	Repair	*0.09677*		rs2287321	1	*0.0839*		rs2287321	1	0.2781
3	rs532411	POLQ	Repair	*0.09839*		rs532411	1	0.399		rs532411	1	*0.0465*
17	rs17222691	RAD51C	Repair	*0.03095*		rs7224276	0.997	*0.0951*		rs10515159	1	0.4186
3	rs3218634	POLQ	Repair	*0.04772*		rs650469	1	0.4721		rs3218634	1	*0.0296*

Table 5.3: SNPs with a p value less than 0.1 in at least two populations (I)

Chr	dbSNP id	Gene	SNP type	p		Houlston tag	r2	Houlston p		Gallinger tag	r2	Gallinger p
1	rs11807227	TP73/LOC339448	Candidate	0.4941		rs11807227	1	*0.0164*		rs2275824	0.031	*0.0546*
10	rs913119	MGMT	Repair	0.5139		rs4751118	0.978	*0.0428*		rs3750826	0.831	*0.0357*
6	rs3218595	REV3L	Repair	*0.08725*		rs12661704	0.008	*0.0334*		rs3218595	1	*0.5056*
14	rs12587397	RAD51L1	Repair	0.5001		rs12587397	1	*0.072*		rs7151629	0.064	*0.0544*
18	rs8305	POLI	Repair	*0.003347*		rs8305	1	0.5497		rs8305	1	*0.0928*
10	rs7085679	DCLRE1C	Repair	*0.06088*		rs10159823	0.996	0.5789		rs3814171	0.139	*0.0506*
1	rs4648414	TP73/KIAA0562	Repair	*0.05095*		rs6683156	0.997	0.6058		rs2275824	1	*0.0546*
18	rs2606246	ENOSF1	Candidate	*0.06502*		rs2606246	1	*0.0748*		rs2606246	1	0.5744
11	rs2251075	C11orf30	Repair	*0.03232*		rs4945087	0.934	0.6323		rs2155220	0.935	*0.0789*
23	rs2301188	CETN2	Repair	0.6673		rs2301188	1	*0.0651*		rs2301188	1	*0.0364*
12	rs6560891	P2RX2	Repair	0.6645		rs10870483	0.896	*0.02*		rs7966242,rs4883616,rs1132375	0.746	*0.088*
22	rs5757133	DMC1	Repair	*0.01885*		rs5757213	0.939	*0.0651*		rs3180098	0.926	0.6961
1	rs1010447	FRAP1	Repair	*0.007617*		rs11121704	1	0.76		rs1010447	1	*0.0131*
1	rs3765761	TP73	Repair	*0.08278*		rs3765761	1	0.6315		rs10492942	0.017	*0.067*
11	rs4944051	POLD3	Repair	0.7831		rs4145953	0.998	*0.0106*		rs4944051	1	*0.0158*
21	rs218634	CHAF1B	Repair	0.7508		rs218634	1	*0.0007*		rs2244034	0.278	*0.0631*
11	rs7939226	POLD3	Repair	0.8004		rs7939226	1	*0.0098*		rs7939226	1	*0.0427*
8	rs1265116	RRM2B	Repair	*0.02591*		rs2925790	0.988	*0.0235*		rs1265116	1	0.8194
1	rs10864490	FRAP1	Repair	*0.0111*		rs1417131	0.983	0.8708		rs6701524	1	*0.0254*
11	rs7943085	POLD3	Repair	0.7988		rs6592577	0.999	*0.0306*		rs7943085	1	*0.0936*
5	rs334888	UNG2	Repair	*0.06878*		rs334888	1	0.8206		rs336081,rs231622,rs389075	0.744	*0.0627*
2	rs335128	WDR33	Repair	0.8913		rs335128	1	*0.0979*		rs335128	1	*0.0022*
15	rs3087374	POLG	Repair	*0.07786*		rs3087374	1	0.9038		rs1599857	0.173	*0.0372*
17	rs4564632	FLJ35220	Repair	*0.05366*		rs6565681	0.335	0.9103		rs4564632	1	*0.0597*
19	rs8101626	DNMT1	Repair	*0.02633*		rs2162560	0.981	*0.0919*		rs1051738	0.035	0.9288
11	rs7944514	POLD3	Repair	0.9406		rs10899013	0.989	*0.0234*		rs7944514	1	*0.0957*
8	rs1208	NAT2	Candidate	*0.0823*		rs1208	1	0.9516		rs1208	1	*0.0399*
8	rs2607661	EDD	Repair	*0.01746*		rs2978431	0.961	*0.0895*		rs2607661	1	0.9971

Table 5.4: SNPs with a p value less than 0.1 in at least two populations (II)

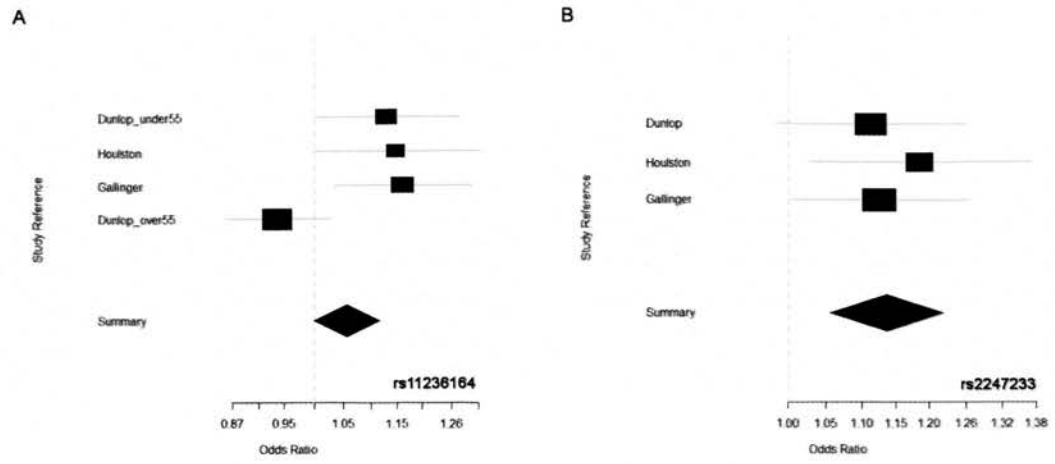


Figure 5.8: Odds ratios for SNPs rs11236164 and rs2247233 in the Scottish population, as well as those of their corresponding proxies in the London and Canadian datasets.
Both SNPs were typed in the Dunlop under 55 cohort initially however rs11236164 had also previously been genotyped in a second Dunlop cohort of older cases and controls and these results were therefore included.

will have such small effects [209]. To overcome this, we are typing our strongest candidates in a further 4000 individuals. Combined with the 2000 Canadian individuals this should give us sufficient power to detect more marginal effects.

5.3.5 Tagging efficiency

It has been proposed from the examination of ENCODE data that picking HapMap phase I tagging SNPs with an r^2 threshold set at 0.8 should capture 75% of all common variants in the human genome. Consequently, in association studies, only a sub-set of all SNPs are selected to be genotyped based on the assumption that typing SNPs in high LD will lead to unnecessary redundancy. This however is only true if the p-value of the tag SNP is, as or more significant than the SNP it is meant to be tagging. If a disease variant that is typed directly has a significant p value but its proxies do not, then picking tagging SNPs may lead to this disease locus being missed. We therefore tested the relationship between the r^2 values observed between SNPs typed in our study and their corresponding p values. To achieve this we calculated the r^2 between every pair of SNPs within 500kb (using the HapMap CEU data) and then calculated the difference between the observed p values at each

of these SNPs. Those pairs of SNPs with an r^2 of approximately 0.8 differed in their p values by on average 0.175. This is of course a large enough value for a disease variant to not be detected through a tag. However a large difference in p values between a pair of SNPs is only important if one of the polymorphisms is significant, if large differences are generally only associated with polymorphisms of large p values then they are of little concern. We therefore excluded all SNP pairs in which one of the SNPs did not have a p value of less than 0.1. The average difference in p values of the remaining SNPs whose r^2 was between 0.8 and 0.9 was 0.065. Of those pairs of polymorphisms (with an r^2 greater than 0.8) where one of the SNPs had a p value less than 0.05, in over a third of cases the p value of the second SNP was greater than 0.05.

However perhaps more important than the relationship between the actual p values observed between pairs of SNPs and their level of LD, is the relationship between the r^2 between pairs of SNPs and their rank among all SNPs tested. Of those SNPs that had at least one proxy with an r^2 of at least 0.8, and whose p value were in the top 5% of all SNPs tested, approximately a third (31%) had a corresponding tag outside the top 5%. For example the tag of the second most significant SNP in the whole study, whose p value was 0.004, was not even in the top 10% of all SNPs (r^2 between these SNPs was 0.87, p of tag was 0.16). Similar results were observed using the Canadian data. Those studies therefore that simply select the top SNPs for further study irrespective of their p value (as done in our whole genome scan) may still miss a substantial number of potential disease loci.

From these results we are able to determine what proportion of disease variants we are able to detect through an association study such as ours. If we make a naive assumption that there are 100 common disease variants within the human genome whose p value would be at most 0.05 in a study of our size, we know that approximately 25 of these polymorphisms will be missed simply due to the incomplete nature of the HapMap data but that, on the other hand, 26 will be typed directly due to having no suitable proxy (Figure 5.3). Tagging the remaining SNPs will lead to approximately 16 disease variants being selected as tags by chance, and of the remaining 33, 11 will not be detected at a significance threshold of 0.05. Consequently, of the initial 100 disease variants, 36 will not be detected through an association study such as ours. Although it may be possible to improve this figure by biasing tag selection towards those SNPs that are more likely to be associated with disease, for example by picking non-synonymous SNPs as tags where possible,

most association studies are still likely to miss a substantial proportion of disease variants.

One of the most likely factors leading to this relatively poor relationship between r^2 and p values is the HapMap data itself. Not only are the r^2 values based on the genotypes from relatively few individuals, but the HapMap Caucasian dataset is likely to differ at least to some extent to the Scottish population. Calculating r^2 values based on the Canadian dataset, that contains genotypes for over 2000 individuals and that we have shown is highly similar to our population, does appear to improve the relationship between the r^2 and p values of pairs of polymorphisms. Although allele frequency is also likely to explain some of the discrepancy between r^2 and p values (SNPs of low minor allele frequency are more likely to have higher r^2) the majority of rare SNPs were excluded from our study (only rare non-synonymous polymorphisms were retained).

5.3.6 Inter-chromosomal LD

It has been argued that when genes interact epistatically, evolutionary selection will promote their genetic linkage as a means of enhancing the coinheritance of favourable allelic combinations. This hypothesis has been supported by the observations of clusters of genes in eukaryotic genomes whose corresponding protein interact and/or are coexpressed [210]. Although it was initially proposed that this may simply reflect underlying gene duplications, it was shown by Pal and Hurst [211] that areas of the genome containing clusters of essential genes display relatively low levels of recombination. Consequently there is support for the hypothesis, first proposed by Nei in 1967 [212], that high levels of linkage can be observed between interacting genes, so that alleles that work well together will be maintained.

This hypothesis was further supported by observations in inbred mice of the coinheritance of optimal sets of alleles among linked genes; recombination that broke up these sets of alleles reduced the fitness of the corresponding mice and their ancestors [213]. In this study LD was also observed between markers on different chromosomes. It has been previously observed that LD between unlinked (different chromosome) markers in the human genome is higher than would be expected by chance (unpublished communication), and we were therefore keen to test if these observations were related. Do functionally related genes on separate human chromosomes display higher levels of linkage than would be expected by chance? For example embryonic

lethal combinations of genotypes between SNPs should not be observed, likewise combinations that affect fertilisation success should be selected against. To test this hypothesis we obtained 1030 interacting gene pairs from the EBI intact and MIPs databases.

To identify whether genes whose products are known to interact display higher levels of LD than we would expect by chance, we first calculated the LD observed between interacting gene pairs located on different chromosomes. These results were then compared to those obtained when the person ids were permuted in one of the genes in each gene pair. This consequently provided us with an indication as to the number of SNP comparisons we would expect to see above a certain r^2 cutoff by chance, given SNPs with the same alleles and minor allele frequencies.

Less than 4% of permutations based on the Chinese HapMap data displayed more pairwise comparisons with an r^2 of 0.4 than the real unpermuted data, with none displaying more with an r^2 greater than 0.5. Likewise only 6% of permutations displayed more pairwise comparisons with an r^2 of 0.4 when the HapMap Caucasian data was used. These results therefore appear to suggest that genes whose protein products interact may indeed display higher levels of LD between their SNPs than we would expect by chance. However when r^2 values were calculated using the Yoruba HapMap data over 19% of permutations displayed more pairwise comparisons with an r^2 of 0.4 than the real data (though this did drop to 8% at an r^2 of 0.5). Likewise the figure at an r^2 of 0.4 using the Japanese data was even higher at 31%. The Chinese and Japanese populations are very similar, especially in their LD structures, and therefore if we believe epistatically interacting genes do indeed show higher levels of LD than we would expect by chance we would have expected the permutation results to have been closer. However as we had observed potentially promising results in at least two populations considering the limitations of the interaction dataset (confidently identifying proteins that interact directly in humans can be difficult) we decided to investigate the potential relationship between epistatically interacting genes further. Although we had shown that genes whose products interact potentially display higher levels of LD than we would expect by chance, we had not shown that these levels of LD were any higher than we would expect between randomly selected pairs of genes. For example, all pairs of genes may show higher levels of LD than would be expected. However when random pairs of genes were selected they generally showed at least ten times fewer SNP pairs with an r^2 of 0.4 than the genes known to interact, and fifteen times fewer with an r^2 of 0.6 (Caucasian genotype

data). It should be noted however that the majority of SNPs with an r^2 of 0.4 are from only a subset of genes whose products are believed to interact. For example, although 65 SNP pairs display an r^2 of 0.4 or greater these are all located in only 10 genes. Consequently rather than all epistatically interacting genes displaying higher levels of linkage than would be expected by chance only a small proportion do and consequently it is not possible to say that the high levels of LD observed in certain populations are a result of the interaction between the genes.

One of the major limitations of using HapMap data for this type of analysis is the relatively few individuals available. There were however over thirty times as many people typed in our custom analysis than in the largest HapMap populations. Likewise, as we had examined a candidate pathway, this set of genes was likely to be enriched with those that interact epistatically. We consequently calculated the r^2 between all pairs of SNPs, located on different chromosomes, genotyped on our custom array. After 100 permutations[2] similar to those discussed above, only one pair of SNPs remained significant (no permutations displayed a higher r^2 value at any pair of SNPs). This pair of non-synonymous SNPs were located in *LIG4* and *ATM*. This pair of genes have previously been shown to interact epistatically and are also known to be associated with embryonic lethality [214]. r^2 may not however be the most appropriate measure for this type of analysis, primarily as it is based on predictions of phase between SNPs. When comparing SNPs on different chromosomes phase, i.e. the parent of origin of each allele, is hard if not impossible to determine. We therefore applied a simple chi square test to the observed frequencies of genotypes between any pair of SNPs and those we would expect given random associations. After permutation analysis the same pair of non-synonymous SNPs did not remain significant array-wide. It is possible that any non-random association between polymorphisms is simply the result of genome assembly errors or SNP mis-annotation and therefore more analysis is required to try and understand why some gene pairs show higher levels of linkage than would be expected by chance.

[2]Only 100 permutations could be run due to to the time required to run each permutation

Chapter 6

SNP Prioritisation

6.1 Introduction

As highlighted in the previous chapter most association studies are by their nature indirect. Disease variants are most often identified through their linkage disequilibrium with a second, tag, variant. But what if it was possible to type disease variants directly? Although this is impossible without knowing the location of each disease variant, it is possible to make predictions as to which SNPs are *potentially* deleterious. For example, coding SNPs make up the majority of disease alleles in Mendelian disorders. Botstein et al. illustrated that of all mutations underlying disease phenotypes, 58.9% were missense or nonsense SNPs [192] despite coding SNPs making up only a small proportion of known polymorphisms (approximately 1.5%). By typing only coding variants it is therefore possible to significantly enrich for potential disease SNPs. Even among coding SNPs many polymorphisms are more likely to be deleterious than others and it is possible to prioritise those with strong evidence of being potentially pathogenic. Although there is likely to be ascertainment bias in the analysis of Botstein et al. (as most investigators preferentially look at non-synonymous polymorphisms there is likely to be an artificially high proportion of non-synonymous polymorphisms associated with disease), and complex diseases like cancer are likely to differ from Mendelian diseases, it is still highly likely that coding SNPs will be overrepresented among cancer disease variants.

All SNPs are not however functional and determining which are is not simple, especially when their contribution to the disease may be small or when various SNPs are in LD. However the polymorphisms most easily identified as potentially pathogenic

are those that can be seen to dramatically affect the structure of a protein. For example nonsense mutations often lead to a considerable truncation of the protein, consequently it is relatively easy to hypothesise that the protein's function is being affected. Missense mutations on the other hand are less easily annotated as potentially pathogenic; as not all amino acids in a protein are equally structurally or functionally important. It has been shown however that changes at positions in proteins that are conserved across species and multigene families are more likely to be detrimental. Programs that utilise conservation to predict deleterious mutations/alleles include SIFT [66], PolyPhen [215] and SNPs3d [207] that are discussed below. Likewise changes between amino acids that are dissimilar in terms of their properties (charge, side-chain length etc) are more likely to be potentially deleterious. Wang et al. illustrated that 90% of pathogenic, missense mutations can be associated with a predicted structural change, compared to only 30% of general polymorphisms [216]. Consequently, if a known structure is available, particular structural consequences can be looked for. Programs that determine potential structural consequences of amino acid changes include SNPs3d and PolyPhen.

Methods for identifying single nucleotide polymorphisms that affect splicing or the expression of a gene have been less well characterised, partly due to the ambiguity in splicing and regulatory motifs. However Clifford et al. showed that missense mutations often adversely affect the score of Pfam domain predictions, and that the magnitude of the change in the score is a good predictor of whether the mutation is deleterious [217]. It may therefore be possible to apply a similar technique on splicing and regulatory domains and programs such as FASTSNP [67] have begun to adopt this approach using ESE and transcription factor motifs.

There has already been some notable successes of the use of SNP prioritisation in the design of association studies. For example Begovich et al. [196] typed 87 polymorphisms in a study investigating rheumatoid arthritis that were selected not on their tagging abilities but because they were located in candidate genes/regions and were putatively functional. A study therefore that includes a combination of both tagging and functionally important SNPs may provide the most power for identifying disease variants.

The aim of this project was to develop an integrated analysis environment for SNP prioritisation that could be used in both the design and analysis of both candidate and whole genome association studies. This would be achieved partly through the integration of a number of third party SNP prioritisation programs and the pre-

sentation of SNP prioritisation results across the human genome. Below I discuss the programs used during this project.

6.1.1 SNP Prioritisation Programs

6.1.1.1 SIFT

SIFT (Sort Intolerant From Tolerant) is a sequence homology-based program that uses multiple sequence alignments to predict amino acid changes that will have a phenotypic affect. It achieves this by first compiling an alignment of sequences using the PSI-BLAST program. The database we used was the June 2006 release of the SWISS-PROT/TrEMBL protein sequence database. Sequences from this PSI-BLAST search that are more than 90% identical to one another are collapsed by SIFT to a single consensus sequence with each sequence position represented by the amino acid that occurs most frequently. PSI-BLAST is then again used to search among these consensus sequences to find the top hit to the initial query sequence. These sequences are subsequently aligned, conservation at each position in the alignment calculated (see algorithm 4) and the median of these values determined. If this value of median conservation is greater than a user defined cut-off, the sequence is maintained in the alignment and a PSI-BLAST checkpoint file created. This file is used to query the remainder of the consensus sequences and the process is repeated until the sequence added falls below the median conservation cut-off. To prevent contamination of the alignment with pseudogenes (or even the query sequence itself), sequences with greater than 90% identity to the query sequence are excluded [218, 66].

Algorithm 4 Measure of conservation used by Ng and Henikoff.

$$R_c = log_2 20 - \Sigma_{20aa} p_{ca} log p_{ca}$$

$p_{ca}$is the frequency at which amino acid a appears at position c. Possible values for median conservation range from 4.3 (sequences are 100% identical) to 0 (all amino acids are found at almost all positions)

The prediction of an individual amino acid's affect on phenotype is a weighted average of the observed frequencies of the amino acid and the Dirichlet estimation. Dirichlet mixtures allow the probability of an amino acid at a position in the sequence to be calculated based on prior information about what types of amino acid distributions are reasonable in columns of alignments [219]. This allows amino acid probabil-

ities to be adjusted according to the amino acids observed at that position and those observed in databases of alignments. For further information on how the probabilities are calculated in SIFT see `http://blocks.fhcrc.org/sift/SIFT_help.html`. Through comparisons to experimental data a probability of 0.05 has been shown to be a reasonable cutoff for calling a change deleterious. SIFT also returns the median conservation of the sequences in the alignment as predictions based on alignments with high median conservation (>3.25) are thought unreliable. When tested on actual SNPs annotated as associated with a disease in the SWISS-PROT database (and not simply SNPs in genes associated with a disease), SIFT predicted 69% were deleterious (i.e. a false-negative rate of 31%). Analysis of random SNPs from the dbSNP database on the other hand resulted in 25% being predicted as deleterious, however a proportion of these are likely to be pathogenic in some way [218].

6.1.1.2 Polyphen

The PolyPhen (POLYmorphism PHENotyping) program uses three sources of information to make its prediction of whether an amino acid change will be deleterious. The first is whether the amino acid is located in a site annotated in the SWALL database as functionally important. Features classified as important by PolyPhen include active, binding and transmembrane sites, and any missense change at one of these sites is predicted to be deleterious. The second source of data used by PolyPhen is structural information. By comparing the query sequence to a database of proteins with known structure PolyPhen attempts to identify whether the query itself or a homologue with at least 50% identity has a known structure. If this is the case, Polyphen makes predictions as to whether the amino acid change will affect the hydrophobic core of the protein, electrostatic interactions, interactions with ligands or other important features. This is achieved by calculating several structural parameters and assessing the impact of the amino acids change on spatial contacts. For example, changes that affect the residue's side chain volume or that disrupt an interaction between subunits of a protein are predicted to be deleterious [215].

The final source of information used by PolyPhen is sequence conservation. In a similar fashion to that of SIFT, PolyPhen uses BLAST to identify homologues of the query sequence that display 30-94% identity over at least 50 base pairs. A profile matrix is then calculated from this alignment using the PSIC software. Each element of this matrix is a logarithmic ratio of the estimated probability of a given

amino acid occurring at a site, given infinitely long evolution, against the background frequency of that amino acid (derived from a database of proteins, see PSIC paper for more details [220]). Polyphen then uses the absolute difference of these profile scores to predict the affect of an amino acid change at a particular position. Like SIFT the confidence of these predictions are determined by the quality of the original alignment.

The performance of PolyPhen on known deleterious variants has been tested on sets of disease causing (as annotated by Swissprot) and control variants (between-species substitutions, fixed differences are thought unlikely to be pathogenic, no numbers were reported). When only variants with strong evidence of being pathogenic are examined, PolyPhen displays a false negative rate of 43% and a false positive rate of 3%. Less strict criteria give rates of 18% and 8% respectively [215].

6.1.1.3 SNPs3d

SNPs3d, like PolyPhen, uses both structural information and sequence conservation to predict the affect of amino acid changes. As briefly discussed previously, Wang et al. illustrated that the majority (90%) of a subset of the known pathogenic missense changes listed in HGMD could be associated with a predicted structural change [216]. A subset of these factors, shown in table 6.1, are used by SNPs3d to predict the structural consequences of an amino acid change. This is achieved by training a Support Vector Machine with both pathogenic and control missense changes, to determine the partitioning surface between these classes of changes in the 15 dimensional parameter space (i.e. the 15 factors in table 6.1).

The prediction of deleterious mutations based on sequence conservation in SNPs3d also relies on a SVM. Like SIFT and PolyPhen an alignment of sequences homologous to the query is built using PSI-BLAST. Where two sequences in the alignment are more than 90% identical one is removed. Sequences with less than 30% identity to the query sequence are also excluded along with regions of the alignment where more than 50% of the sequences have a gap. The SVM is subsequently trained on five features associated with conservation; the probability of accepting the amino acid substitution (derived from the BLAST position specific scoring matrix), the entropy at the position in the alignment, the mean entropy across the whole sequence, the standard deviation of the entropy across all positions and the entropy expressed as a Z score [207].

Type	Factors
Continuous factors	Electrostatic interaction: polar-polar, polar-charge, charge-charge
	Over-packing
	Hydrophobic burial
	Surface accessibility
	Structural rigidity: crystallographic β-factor, Z score and standard deviation
Binary factors	Cavity
	Electrostatic repulsion
	Backbone strain
	Buried charge
	Buried polar
	Breakage of a disulphide bond

Table 6.1: Factors affecting protein stability used by SNPs3d to investigate the affects of missense changes. Taken from Yue, Li and Moult

In both the protein stability and conservation profiles the control variants used to train the SVMs were non-synonymous base differences observed between human sequences and their orthologues in closely-related mammals. The deleterious variants on the other hand, were nonsynonymous polymorphisms previously associated with monogenic diseases (obtained from the Human Gene Mutation Database). Random subsets of each class of variant were used to train and test the SVMs so that false-positive and false-negative rates could be estimated. In SNPs3d, variants with an SVM score greater than 0.5 are annotated as non-deleterious and less than -0.5 as deleterious. Predictions between these scores are deemed unreliable. The false-positive and false-negative rates obtained using these SVMs and parameters were 6% and 16% for the conservation SVM, 12% and 21% for the stability SVM and 3% and 9% for both combined (i.e. where a variant is confidently predicted to be deleterious by both methods). Therefore 6% of the variants annotated by the conservation SVM as deleterious in test datasets were in fact not deleterious, and 16% of the variants not annotated as deleterious were in fact pathogenic. It is hypothesised that the error rates are higher when predictions are based on structural information alone (than when based on sequence conservation or both combined) as many affects on protein function such as post-translational modification will not be detected by this

method. On the other hand the conservation profile should theoretically be able to detect any deleterious affect on a protein's function [207].

These programs were used to prioritise SNPs in and around our genes of interest. A number of points should be made about these programs. The first is that each was tested on a different dataset so that the error rates reported above are not comparable between programs. Likewise, all three of these programs were tested and/or trained on disease variants associated primarily with Mendelian diseases. Analysis of the performance of these programs on complex diseases such as cancer is limited and often contradictory. It is likely that variants associated with cancer are less strongly conserved than those involved in Mendelian diseases, not only because of the complex nature of the disease but also the fact that the disease is generally later in onset. However, this may simply mean that a less stringent cut off is required for a polymorphism to be annotated as potentially pathogenic in cancer and other complex diseases (at the expense of increasing the rate of false positives). Zhu et al. have previously shown a weak but significant correlation ($p = 0.002$; $r^2 = 0.06$) between SIFT tolerance indices for 46 SNPs in human repair associated genes and their corresponding odds ratios derived from cancer epidemiological studies [208]. This correlation was confirmed in a larger dataset in chapter 5. SNPs3d is most likely to be affected by the lack of available data on complex diseases as it, unlike SIFT and PolyPhen, is actually trained on Mendelian diseases (whereas the others have simply had their cutoffs determined from Mendelian disease data). We have therefore tried to use raw data where possible rather than predictions (e.g. benign, deleterious etc) in this study.

6.2 Methods

6.2.1 SNPViewer

The SNPViewer program for prioritising SNPs, was implemented in the Java programing language. A number of third party programs were included in the final implementation of the program:

- EnsJ: An open source Java library providing access to current and archived Ensembl databases. Developed by Ensembl and used in SNPViewer to retrieve gene and SNP information. Now no longer supported by Ensembl.

- MySQL Connector/J: For connecting to MySQL through Java. Required for connection to Ensembl databases via EnsJ.
- Haploview: Haplotype analysis Java program. Code adapted to function within SNPViewer. (adaptions to pass genotypes and to pass clicks)

SNPViewer web accessibility was implemented via Java web start. The class diagram for SNPViewer is shown in figure 6.1 and the program is accessible at `http://www.hgu.mrc.ac.uk/Users/James.Prendergast/PolyAnalyser/SNPViewer.jnlp`.

The graphical user interface of SNPViewer is split into four panels as shown in figure 6.2. The control panel, located on the right of the GUI, contains a text area that displays information on the currently selected SNP, checkboxes for selecting the data to be displayed on the frequency plot and a spinner for selecting the repair gene to view.

The main panel of the SNPViewer interface is the frequency plot. The frequency plot displays minor allele frequencies (MAFs) and average heterozygosity for SNPs where such information is available. This data was collected from four different sources: Ensembl, HapMap, ABI, and NCBI. Data from each source is represented by a different shape:

- Ensembl Minor Allele Frequency : o
- HapMap Minor Allele Frequency : □
- ABI Minor Allele Frequency : x
- NCBI average Heterozygosity : +

Each of these shapes is also colour coded according to the type of SNP it represents (relative to the gene of interest only):

- Flanking SNP : Green
- UTR SNP : Blue
- Coding SNP : Red
- Intronic SNP : Orange

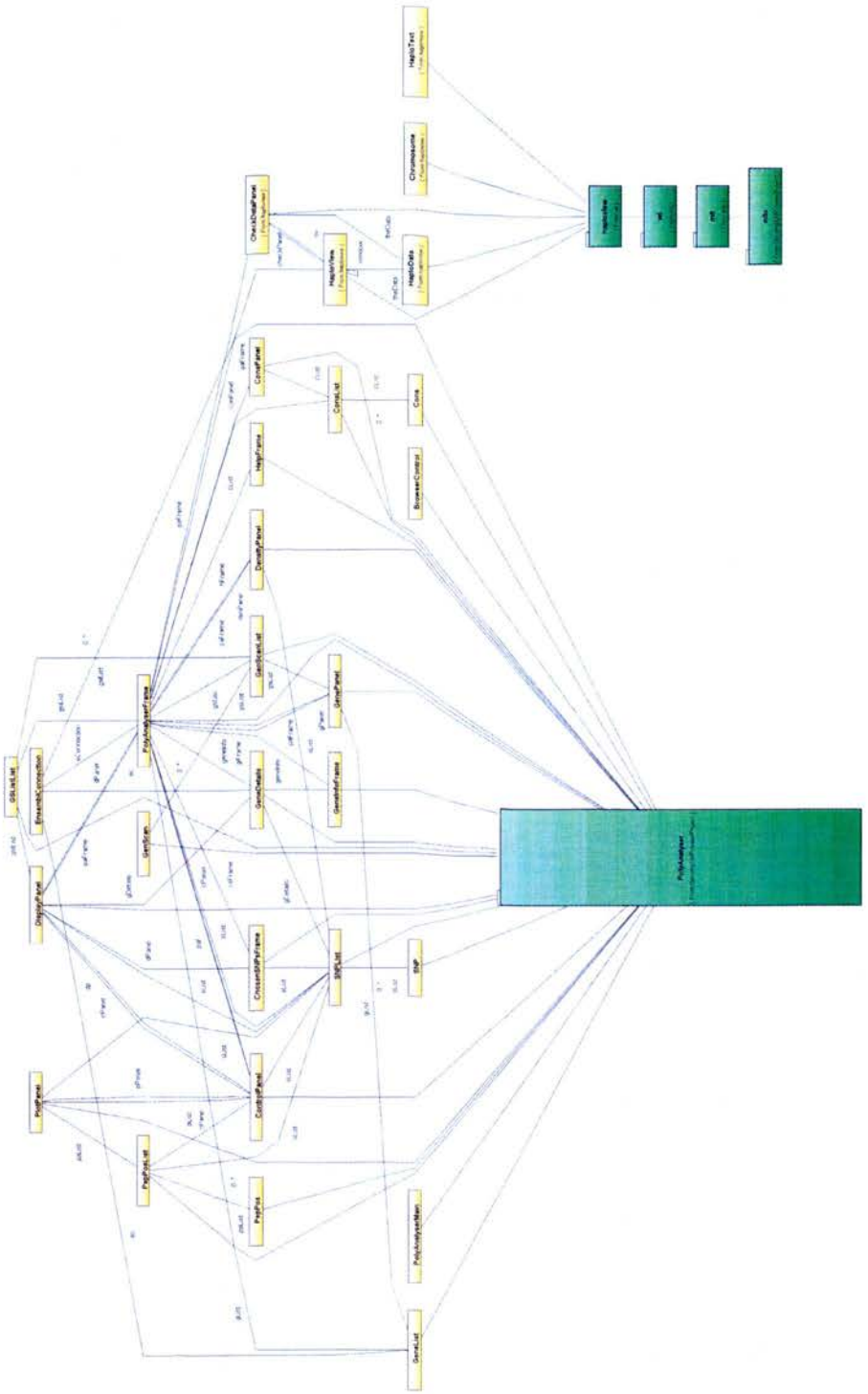


Figure 6.1: SNPViewer class diagram.
Only edited Haploview classes are shown.

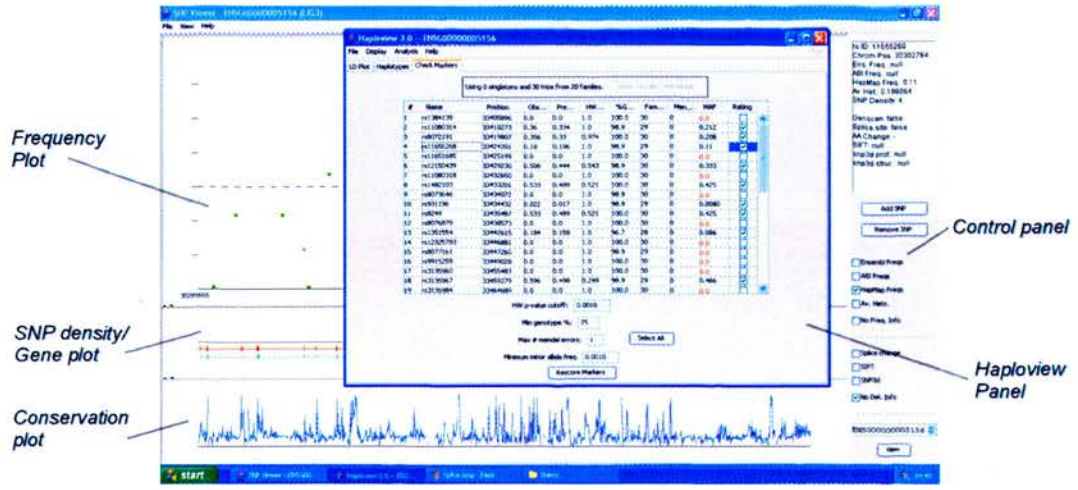


Figure 6.2: Graphical User Interface of SNPViewer

In this plot the x-axis represents base position along the chromosome and exons (as defined by Ensembl) are marked below it in red. Each time the plot is changed the program retrieves the most up to date positions of these exons. The y-axis is either MAF or average heterozygosity (depending on the data point) and is marked by 0.1 increments. The dashed line across the plot is at 0.3.

In order to select a SNP it can simply be clicked in the frequency plot. This will lead to its information being displayed on the control panel where it can subsequently be added to a list of chosen SNPs (or by right clicking). Those SNPs that have already been chosen are coloured in black.

Directly below the Frequency plot is the SNP density plot. Each vertical line in this panel represents a SNP and its height represents the number of known SNPs within 2000bp. Both plots are aligned so that evidence for a particular SNP in the SNP density panel will be directly above it in the Frequency panel. The y-axis increments by 5 SNPs and the x-axis and colour coding are the same as for the Frequency plot. This panel can be changed to display Ensembl genes and GENSCAN predictions instead of SNP frequencies.

The conservation panel is shown at the bottom of the SNPViewer GUI. The initial conservation trace implemented in the first version of SNPViewer was calculated using the sum of pairs method. This involved creating an alignment between the

human protein of interest and all predicted Ensembl orthologues using the t-coffee alignment program. A suitable matrix was then used (in this case PAM120) to score the similarity between all combinations of the residues observed at each position. The sum of these scores consequently provides us with an indication of the level of conservation observed at each amino acid. This initial conservation plot however had two drawbacks. Not only was creating the alignments slow but conservation scores were only available for coding amino acids. Therefore this trace was replaced by the vertebrate multiz alignment from the UCSC genome browser.

6.3 Results and Discussion

The aim in developing SNPViewer was to provide a program that could be used to identify functionally important SNPs to aid in the design and interpretation of our association studies.

As mentioned above most association studies are by their nature indirect. Disease variants are most often identified through their linkage disequilibrium with a second, tag, variant. However we were keen to attempt to improve our association study design by enriching for tags with potential phenotypic consequences. To test the feasibility of using SNP prioritisation in designing association studies SNPViewer was first trialled using the DNA repair genes examined in chapter 5.

Five initial sources of information were used:

1. SNPs3d.
2. SIFT.
3. Polymorphisms affect on GENSCAN prediction.
4. Polymorphism causes nonsense change (on Ensembl gene prediction).
5. Polymorphism in conserved region.

GENSCAN was written by Chris Burge and is a program designed to predict genes and gene structures in DNA sequences [221, 222]. On vertebrate sequences containing short genes with simple exon structures GENSCAN has been shown to have a sensitivity and specificity per nucleotide of 0.93[1]. In SNPViewer a user is able to

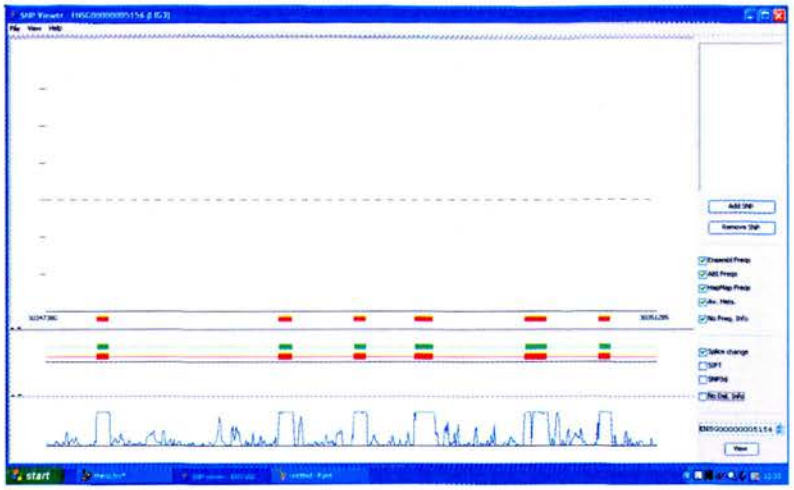
[1]Sensitivity and specificity was calculated by comparing GENSCAN predictions to a set of known genes.

compare GENSCAN predictions of DNA sequences containing the various alleles of a polymorphism. If predictions differ depending on a SNPs allele we predict this variant affects the mRNA structure of a gene. GENSCAN prediction can however be inaccurate and we therefore also compare the predictions to predicted gene structures from Ensembl. Although Ensembl gene structures are partly based on GENSCAN results we treat those exon predictions that match Ensembl genes as high quality. If therefore a GENSCAN prediction that matches an Ensembl gene is affected by an allele change we predict that this SNP is potentially deleterious. An example is shown in figure 6.3.

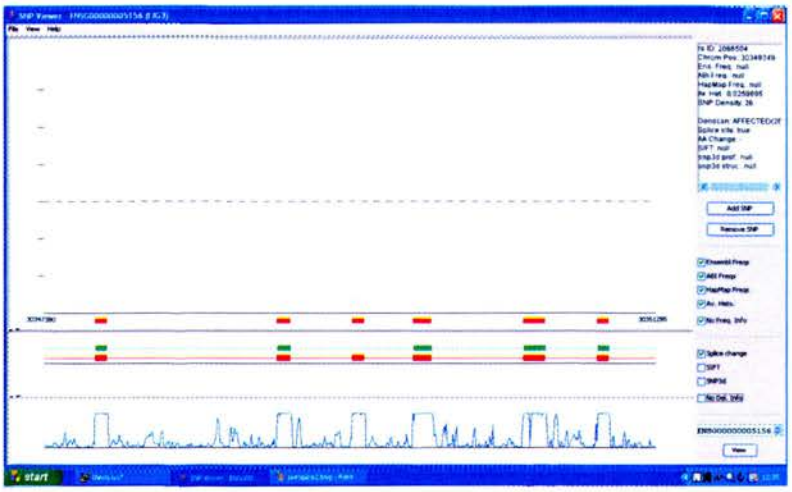
To test the feasibility of using GENSCAN to predict splice changes in silico, the G382D polymorphism of the MUTYH mismatch repair gene, predicted by GENSCAN to affect the mRNA structure of MUTYH in humans, was tested by Dr Susan Farrington in vitro. cDNA analysis confirmed that this polymorphism did indeed affect splicing and although far from a rigorous test this result seemed to confirm the potential of using GENSCAN to predict polymorphisms that affect splicing [223].

As shown in chapter 5 the use of SIFT but not SNPs3d is also likely to be a viable approach for prioritising tagging SNPs. SNPs with a high chi-square test statistic appear to be, in general, well conserved. Although the use of programs such as SIFT is restricted to missense polymorphisms, the conservation of the surrounding region is a potential alternative measure that could be used for all classes of polymorphisms. For example SNPs in key, well conserved motifs or domains are perhaps more likely to be deleterious. However we could observe no relationship between a polymorphisms test statistic and the average conservation of the surrounding region as measured using the UCSC multiz 17 species alignment scores. It may simply be the case that there are simply not enough true positives in our dataset to detect a relationship. There is however no high-quality dataset of known cancer associated polymorphisms available. As shown in chapter 5, few if any SNPs with a previous association to colorectal cancer replicate in other populations.

PolyPhen, that predicts whether a particular polymorphism is likely to be deleterious according to a range of criteria, also displayed no relationship to the chi-square values of our non-synonymous repair SNPs (Kruskal-Wallis $p=0.785$). Consequently although we had typed all known non-synonymous SNPs in a pathway critical to tumour progression only SIFT displayed any substantial evidence of being a viable approach to SNP prioritisation. However even its use is likely to provide at best a modest enrichment. Programs such as SIFT, PolyPhen and SNPViewer may there-



(a) GENSCAN prediction of DNA containing allele1 of rs2066504



(b) GENSCAN prediction of DNA containing allele2 of rs2066504

Figure 6.3: GENSCAN predictions in SNPViewer.
Note the missing exon in the GENSCAN prediction (green rectangles and lines) in figure b. The Ensembl gene structure is shown in red. The rs2066504 polymorphism is therefore predicted to affect splicing.

fore be most useful in interpreting the results of association studies rather than in their design, though it is still to be demonstrated that any programs except SIFT are applicable to complex disorders.

6.3.1 SNP class and conservation

As previously stated it is widely assumed that non-synonymous polymorphisms, as a class, are more likely to be associated with cancer than SNPs in general. Unlike synonymous or intergenic SNPs, non-synonymous polymorphisms can directly affect a protein's structure. In this study, we had therefore typed all validated, non-synonymous SNPs that were located in one of our repair genes of interest.

Although most polymorphisms that are associated with cancer are non-synonymous, this is primarily the result of most studies, like us, preferentially studying non-synonymous SNPs. However the whole genome study gave us the opportunity to investigate whether non-synonymous polymorphisms were indeed more often associated with cancer than other classes of SNPs.

We did this by first determining the class of each of the SNPs in the whole genome study via the Ensembl Perl APIs. Allelic chi-square values for each whole genome SNP were then calculated. Using this data we were able to compare classes of genes and look at whether non-synonymous SNPs, on average, showed a higher chi-square value than other types of SNPs. As can be seen in table 6.2 the average chi-square value of non-synonymous polymorphisms was not significantly different from SNPs in general. In fact the only class of SNPs that showed a significantly higher distribution of chi-square values than the rest of SNPs was regulatory and 5'UTR polymorphisms ($p=0.049$, Mann-Whitney U test, uncorrected for multiple testing). This result would therefore seem to suggest that polymorphisms that control the expression of genes are the most important in cancer with non-synonymous polymorphisms showing no significant difference to intronic and intergenic polymorphisms.

However most polymorphisms in the genome are likely to be neutral with respect to cancer. Any signal from functionally important non-synonymous SNPs is therefore likely to be drowned out by the large number of functionally neutral SNPs. We therefore attempted to enrich for polymorphisms that were likely to be functionally important by examining only those polymorphisms in genes previously associated with cancer (list obtained from the cancer gene census [224]). As shown in figure 6.4 synonymous and intronic polymorphisms located within cancer genes, in general,

SNP type	SNP count	Mean chi-sq
SYNONYMOUS_CODING	2024	0.888
3PRIME_UTR	3129	0.920
INTERGENIC	172175	0.927
INTRONIC	114486	0.931
NON_SYNONYMOUS_CODING	6370	0.942
DOWNSTREAM	7762	0.943
STOP_GAINED	53	0.950
UPSTREAM	8551	0.962
5PRIME_UTR	590	1.056
REGULATORY_REGION	261	1.060

Table 6.2: Mean chi-square by SNP class

show no significant difference in their chi-square values than synonymous and intronic polymorphisms located in other genes. A significant difference is however observed when non-synonymous SNPs are examined. Although strictly the difference is not large enough to survive a Bonferroni multiple test correction (a p value of 0.017 would be required) these data are based on only one study and the addition of results from further studies and cancer types may improve this analysis. There is also likely to be some bias in this analysis as many cancer genes will have been annotated as involved in cancer through a non-synonymous polymorphism. However the extent of this result is perhaps surprising. We would expect linkage disequilibrium between polymorphisms to break down the association between a SNPs type and its association with cancer. For example, we would expect many synonymous and intronic SNPs to be in strong LD with non-synonymous polymorphisms nearby (as the non-synonymous SNPs were added as additional content and not as tags). If this was the case we may expect some difference between the synonymous polymorphisms in the cancer genes and those in the whole genome (in the same way as the non-synonymous polymorphisms) however this is not observed.

There is therefore at least some evidence that SNP prioritisation may be a viable approach to association study design and interpretation. The test statistic of regulatory SNPs is on average 14% higher than intronic and intergenic SNPs. Likewise, when only genes previously associated with cancer are examined, non-synonymous SNPs display on average a 35% higher test statistic than intronic SNPs (and regulatory SNPs 70%, though when based on limited numbers). Perhaps unsurprisingly therefore the best approach to designing cancer association studies is to preferentially

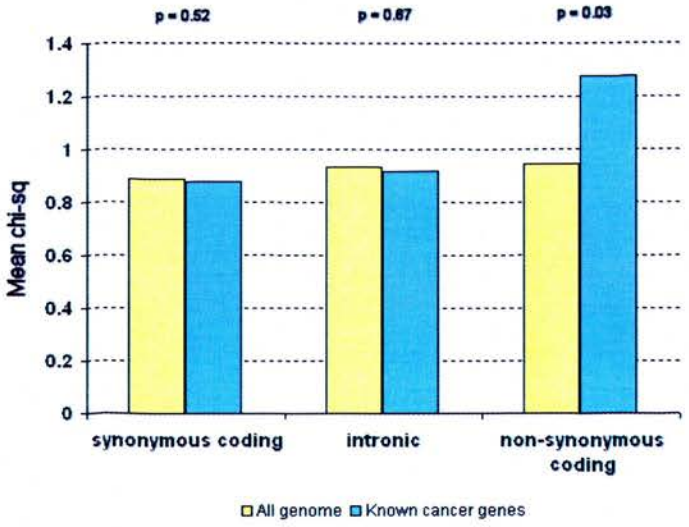


Figure 6.4: Mean chi-squares, by SNP type, for polymorphisms in both cancer genes and the entire genome.
Mann-Whitney values comparing the medians of each group are also shown. (Only ten regulatory SNPs were present in this dataset; genome mean chi-sq: 1.06, cancer genes mean chi-sq: 1.56; Mann-Whitney p=0.12)

select regulatory and non-synonymous tags. Where more than one non-synonymous SNP is in high LD SIFT could then be used to identify the strongest tag. No further methods of SNP prioritisation examined showed any evidence of being applicable to cancer association studies.

To make SNPViewer fully functional it would require a number of additions and alterations. Firstly most results are pre-computed. The time required to make SIFT, SNPs3d and GENSCAN predictions makes it unfeasible for predictions to be made *ab initio*. This does however mean the displayed results are restricted to known SNPs, and therefore a final version of SNPViewer would require the ability for the user to input novel polymorphisms for prioritisation. The implementation of this ability could be achieved by setting up SIFT, SNPs3d and GENSCAN Web Services that could be queried from the Java application. However this is implemented, retrieval of results would be slow and it would therefore be necessary that the query occurred in the background and results could be saved.

SNPViewer is also currently gene-centric. It would however be simple for SNPViewer to be made analogous to a genome browser so that the user could enter regions to be viewed. This would simply require the predictions to be stored in a database

that was queried from the Java application. At present there is the ability to open a genes Ensembl GeneView page from the application, this would have to also be changed to the corresponding Ensembl ContigView page.

We would also add more functionality to the program. For example we could determine the affect of polymorphisms on motif scores. As mentioned previously Clifford et al. showed that missense mutations often adversely affect the score of Pfam domain predictions, and that the magnitude of the change in the score is a good predictor of whether the mutation is deleterious. The impact of polymorphisms, on domain, transcription factor and splice enhancer motifs could therefore also be displayed by SNPViewer.

The wrong approach was however probably adopted in the design of SNPViewer. At present the third party, open source Haploview program was integrated into a novel Java browser. However a far simpler approach would have been to adapt the Haploview program alone. The core of association studies is LD structure that is well handled by Haploview. Likewise the design of Haploview allows new tracks to be added relatively easily, so that SNP prioritisation results could be simply presented. By adapting Haploview to work via Web Start and to read information from a remote database (that we have already implemented) Haploview could have the same functionality of SNPViewer in a far more integrated package. This is therefore the route forward for this project.

Chapter 7

Whole Genome Association Study

7.1 Introduction

In chapter five a candidate gene association study was discussed, however this is only one approach to identifying potential disease genes through the use of association studies. An alternative is to scan the whole genome for potential disease variants rather than simply a number of key candidate regions. The major difference between these two approaches is that a candidate gene association study is hypothesis driven, we are only testing regions of the genome we believe likely to be associated with the disease, whereas no such specific hypothesis is being tested in a whole genome study. Each approach has distinct advantages and disadvantages. A major advantage of candidate region association studies is that it would be hoped that disease variants (or SNPs in LD with disease variants) are enriched among the set of markers being tested so that the number of true positives per SNP tested will be higher (though this of course is not guaranteed). Likewise it should be easier to detect these true positives as the background noise of false positives should be lower, and therefore disease variants of more marginal affects may be detected. Candidate region association studies are also simply cheaper, with large whole genome scans being prohibitively expensive (although the cost per SNP is less). Whole genome studies are not however limited to only detecting associations between a disease and regions with some previous evidence of being associated with the disorder, completely novel associations can be detected. Consequently in the Dunlop lab both approaches were adopted.

In recent months a number of whole genome association studies have been published. One notable success has been the identification of a region on 8q24 by Gud-

mundsson et al. [225] and Yeager et al. [226] that was independently shown to be associated with prostate cancer risk. A third study that had fine mapped this region as a result of prior admixture mapping also found evidence of an association between this region and prostate cancer [227]. In total 21,063 cases and controls from a variety of ethnic backgrounds had been examined across the three studies. Closer examination of this region suggested that a number of loci are involved in prostate cancer risk in this region, as multiple association signals were detected that were separated by regions of high recombination [228]. However none of the polymorphisms displaying the strongest association with the disease are located in known genes and the nearest candidate gene, MYC, is approximately 260kb telomeric [228]. Consequently it is unlikely this region would have been detected by the use of candidate gene studies.

Subsequent examination of a single SNP in this region in breast cancer did not replicate the results observed in prostate, leading the authors to hypothesise that this region does not harbor a region associated with all types of hormone-related cancer [229]. However, even the association between this region and prostate cancer is not straightforward. It appears that at least three independent regions at this locus modify cancer risk, with little linkage disequilibrium observed between them. The combined affect of these regions appear to follow a multiplicative model [228]. Consequently further examination of this region is required to try and further characterise the role of this region in cancer.

7.2 Methods

Genotype, person and SNP information data were stored in three MySQL tables. The Haploview source code was adapted to calculate single and multi-marker LD values between SNPs/haplotypes. Missing genotypes were predicted using the IMPUTE program [209]. IMPUTE works by predicting the probability of observing each possible genotype at each unknown polymorphism given the observed alleles surrounding it and the known haplotype structure and recombination rates of the region (both of the latter obtained from HapMap). Due to the uncertainty in genotype calls arising from IMPUTE we used the SNPTEST program to analyse the resulting data. SNPTEST simply converts those genotypes with a maximum posterior probability greater than 0.9 to the respective genotype and leaves the rest

uncalled (i.e. null) and then performs standard frequentist association tests (additive, dominant, recessive, general and heterozygote models) [230, 209]. The Illumina Beadstudio software was used for the examination of copy number changes. Case control matching and Hardy-Weinberg equilibrium analysis were performed by Dr Tenesa. Resequencing was carried out by Dr Farrington.

7.3 Results and Discussion

The whole genome association study being carried out within the Dunlop lab will ultimately consist of three phases. Phase I, that has already been completed, involved the typing of 555,512 SNPs in approximately 1,000 cases and 1,000 age and sex matched controls. Only cases aged under 55 were typed in this phase in an attempt to enrich for individuals with a larger genetic contribution to their disease. Phases II and III will involve the typing of the most significant polymorphisms from each previous phase in further cohorts of individuals. In total the most significant 1,000 SNPs should be typed in approximately 8,000 Scottish individuals. Although currently only the first two phases of this study have been completed, the availability of genome-wide case control data on a large number of individuals did provide us with the opportunity to collaborate with other groups and combine data.

This collaboration was initially led by the Gallinger group in Canada. Through their own two stage study consisting of typing 99,632 SNPs in 1226 cases and 1239 controls followed by typing the best 1,143 of these polymorphisms in a further cohort of 1,139 cases and 1,055 controls, they had identified 76 SNPs potentially associated with colorectal cancer (see [231] for more details). To determine which of these polymorphisms were true risk variants we identified which of these SNPs had been typed in our own whole genome study (the Gallinger group had been using Affymetrix SNP arrays and consequently there was only limited direct overlap between our two studies). Of the 76 Canadian variants 28 were present on either the 317k or 240k Illumina array and had consequently been genotyped directly in our study in phase I. Through the adaptation of the tagger algorithm implemented in Haploview we identified the best single or multi-marker tag for each of the remaining SNPs. Allelic chi-square p values were subsequently calculated for each of these tags/variants in our dataset. The results can be seen in tables 7.1 and 7.2. The best tag of 7 of the 76 variants was in only limited linkage disequilibrium ($r^2 < 0.7$) and consequently these

variants were typed directly in our cohort by Dr Farrington. Of the 76 Canadian polymorphisms 9 (or their corresponding tag) had a p value < 0.1 in our Scottish population. These 9 polymorphisms were subsequently genotyped in our phase 2 cohort of 1,910 cases and 1,985 controls. After this phase only two variants remained significant (allelic $p < 0.05$). The London dataset was not used in this study due to a potential conflict of interests with their own publication [232].

The first of these variants, rs719725, is located on chromosome 9 approximately 34kb downstream from the nearest known gene, the protein kinase *TPD52L3*. *TPD52L3* is a paralogue of the tumour protein *TPD52*, displaying approximately 26% conservation at the protein level. *TPD52* was initially identified in 1995 by Byrne et al. due to its overexpression in breast carcinomas [233] and is believed to be involved in calcium-mediated signal transduction and cell proliferation [234].

However the nearest gene downstream of this locus, the E3 ubiquitin ligase *NIRF* (rs719725 is located ~50kb 5' of this gene), has also been associated with cancer. *NIRF* is expressed abundantly during cell proliferation but suppressed during cell arrest. Likewise tumour cells have been shown to constitutively express NIRF, and consequently NIRF has been associated with having a role in cell cycle progression [235]. The rs719725 polymorphism is in LD with both of these loci, likewise rs719725 is situated only approximately 15kb from a DNAse I hypersensitive region [236], and only 20kb from a CCCTC-binding factor region [237], suggesting regulatory sites may also be located in close proximity.

Further examination of this locus in French and European cohorts of individuals by the Canadian consortium (a total of 2199 cases and 2401 controls) did not support its association with colorectal cancer. However as copy number changes at this region of chromosome 19 have previously been associated with cancer in a number of studies (e.g. [238, 239, 240]) this locus has not been completely disregarded.

The second polymorphism, rs10505477, is located on chromosome 8q in the intron of a gene of unknown function that was first identified last year as a result of an investigation into an association between this locus and prostate cancer [241]. A second gene, *POU5F1L1*, is also located within 20kb. At present little is known as to the function of either of these genes however *POU5F1L1* appears to have arisen through retrotransposition relatively recently in the primate lineage [242]. Its parent gene, the transcription factor *POU5F1*, may play a role in tissue differentiation and early mammalian development [243]. (Multiple test corrections, haplotype and epistasis analysis and Hardy-Weinberg equilibrium checks at these loci were carried

Arctic SNP	Best tag	r2	Arctic p	p (of tag)	p (of Arctic SNP)
rs10489565	rs3795357	*0.656*	0.0044	0.0270	0.0220
rs10516168	rs12186237,rs13130037 12	*0.433*	0.0016	0.5592	0.0370
rs10493889	rs12024594	0.73	0.0008	0.0213	0.0410
rs11236164	rs3902018,rs10219203,rs7932922 242	0.967	0.0299	0.0544	0.0436
rs10491268	rs1422446,rs11742783,rs247205 421	1	0.0072	0.0629	0.0799
rs3743262	rs7162336	*0.285*	0.0953	0.2938	0.1089
rs10512472	rs9897552	1	0.0027	0.2647	0.2488
rs10769224	rs10742787	1	0.0097	0.2677	0.2797
rs4931434	rs10843881,rs4031375 11	*0.502*	0.0730	0.7458	0.2917
rs1454027	rs1869472	1	0.0036	0.3646	0.3137
rs10484791	rs1523932,rs7746943,rs2204285 243	1	0.0057	0.4279	0.3706
rs9328033	rs6596783,rs1997773 24	0.758	0.0082	0.5574	0.3786
rs2853129	rs2458416	0.483	0.0816	0.2966	0.4545
rs10280428	rs2214726,rs7794797 33	0.451	0.0055	0.3182	0.5255
rs2278170	rs11032345	1	0.0079	0.5994	0.5774
rs9898	rs3733159,rs3733011 44	0.884	0.0081	0.4694	0.5814
rs10489525	rs12408865,rs8453,rs2144428 131	0.884	0.0067	0.8401	0.5814
rs945881	rs10747482,rs11165879,rs7517433 132	0.709	0.0275	0.8976	0.6424
rs1402582	rs1081896,rs1402578,rs614673 112	0.994	0.0016	0.6282	0.6533
rs1504175	rs3788982,rs11903014,rs1504188 212	0.867	0.0075	0.7240	0.6723
rs2049064	rs955988	1	0.0020	0.7323	0.6803
rs2320590	rs3767248,rs4654873,rs9426736 134	0.949	0.0037	0.9584	0.6933
rs6601328	rs4841171	1	0.0056	0.8142	0.7183
rs444772	rs145290	1	0.0131	0.5524	0.7393
rs10499162	rs6935162,rs9321172 32	*0.329*	0.0007	0.0242	0.8032
rs7939226	rs4944925,rs4145953,rs10899013 134	1	0.0779	0.9969	0.8541
rs10498243	rs1549567	1	0.0036	0.9461	0.9374
rs812824	rs1081896,rs614673 12	1	0.0058	0.9052	0.9640
rs10483802	rs910315	0.743	0.0067	0.0025	#N/A
rs719725	rs7857628,rs2066362 13	1	0.0046	0.0025	#N/A
rs10505477	rs6983267	0.935	0.0030	0.0310	#N/A
rs850470	rs850476	0.961	0.0069	0.0669	#N/A
rs797206	rs797208	0.951	0.0047	0.1099	#N/A
rs10517602	rs17031957	1	0.0063	0.1375	#N/A
rs2963765	rs751485,rs269511 23	0.967	0.0029	0.1415	#N/A
rs10507308	rs9578469	*0.487*	0.0073	0.1658	#N/A

Table 7.1: Top SNPs from the ARCTIC (Assesment of Risk of Colorectal Tumours In Canadians) whole genome association study with corresponding Dunlop study tags and p values I.
(Arctic p corresponds to the the p value of the ARCTIC SNP in the ARCTIC study, p of tag corresponds to the p value of the best Dunlop tag of the ARCTIC SNP, p of ARCTIC SNP corresponds to the p value of the ARCTIC SNP in the Dunlop study if genotyped.)

Arctic SNP	Best tag	r2	Arctic p	p (of tag)	p (of Arctic SNP)
rs572619	rs568306,rs491111 23	1	0.0063	0.1769	#N/A
rs10512404	rs17204340	*0.308*	0.0061	0.1950	#N/A
rs377685	rs12482714	1	0.0022	0.2158	#N/A
rs2355084	rs7783055,rs7782151 44	1	0.0060	0.2171	#N/A
rs10518098	rs1398982	1	0.0073	0.2288	#N/A
rs1994967	rs1495271,rs1108993 23	0.821	0.0083	0.2308	#N/A
rs946807	rs10867398	*0.536*	0.0053	0.2476	#N/A
rs10494240	rs720899	1	0.0030	0.2607	#N/A
rs4484159	rs4858703	0.825	0.0076	0.2837	#N/A
rs798893	rs103294	0.945	0.0071	0.2957	#N/A
rs10503122	rs677592	1	0.0038	0.3209	#N/A
rs1399176	rs1514203	0.959	0.0007	0.3419	#N/A
rs10505685	rs1514203	1	0.0040	0.3497	#N/A
rs4372639	rs6496067	1	0.0076	0.3636	#N/A
rs4133195	rs12694428,rs10804264,rs4674259 343	0.948	0.0049	0.3858	#N/A
rs1057083	rs4961323	1	0.0995	0.3986	#N/A
rs360659	rs360622	1	0.0057	0.4336	#N/A
rs431319	rs11913109,rs933582,rs6005625 244	1	0.0071	0.4618	#N/A
rs10503262	rs10089026	0.958	0.0033	0.4675	#N/A
rs3864498	rs4725830	*0.052*	0.0017	0.5874	#N/A
rs10512028	rs6560355	1	0.0041	0.6003	#N/A
rs2469583	rs2433363	1	0.0013	0.6214	#N/A
rs4404442	rs4583651	1	0.0031	0.6444	#N/A
rs10497667	rs6434164	1	0.0010	0.6479	#N/A
rs9830734	rs10513799	0.636	0.0098	0.6484	#N/A
rs4944051	rs4145953	1	0.0468	0.6634	#N/A
rs723142	rs796398	0.959	0.0034	0.7253	#N/A
rs1087	rs9316180	0.961	0.0260	0.7443	#N/A
rs26764	rs26762	1	0.0045	0.7512	#N/A
rs508106	rs796398	0.958	0.0016	0.7687	#N/A
rs4941537	rs4942460	1	0.0593	0.8272	#N/A
rs724667	rs3136559	1	0.0039	0.8442	#N/A
rs10510558	rs1872143,rs322706,rs1561115 243	0.999	0.0048	0.8505	#N/A
rs10502694	rs10502692	1	0.0029	0.8671	#N/A
rs1963296	rs11869275,rs4792347,rs2674958 112	1	0.0058	0.8917	#N/A
rs890248	rs977439	1	0.0057	0.9064	#N/A
rs973128	rs975951	1	0.0015	0.9666	#N/A
rs7200548	rs2118014	0.812	0.0030	0.9670	#N/A
rs1988515	Untaggable	0	0.0061	#N/A	#N/A
rs2075322	Untaggable	0	0.0526	#N/A	#N/A

Table 7.2: Top SNPs from the ARCTIC study with corresponding Dunlop tags and p values II (as above)

out by the Canadian ARCTIC consortium and can be viewed in [231]).

This polymorphism on 8q is located in the same region as that associated with modifying risk in prostate cancer. There was therefore a requirement to try and narrow down the region of 8q associated with colorectal tumours as no genes in the region had a strong previous association with cancer (the nearest strong candidate, the oncogene *MYC*, is approximately 340kb telomeric and is separated from the locus by high levels of recombination).

Our first approach was to compare the results from the three cancers examined to date; prostate, breast and colon. The genotyping results of the prostate and breast studies have been made publicly available (`http://cgems.cancer.gov`) allowing us to make a direct comparison between tumour types (the breast and prostate scans, like our colon study, had been carried out using Illumina arrays). To compare datasets we first calculated the odds ratio at each SNP and then combined the corresponding probabilities across datasets using a method developed by Fisher ($-2\Sigma lnP$). Of all the ~500,000 SNPs present in all three studies, the top three most significant polymorphisms were all located on 8q24 (p ~ 10^{-7}). Shown in figure 7.1 are the combined p values for each SNP across this locus.

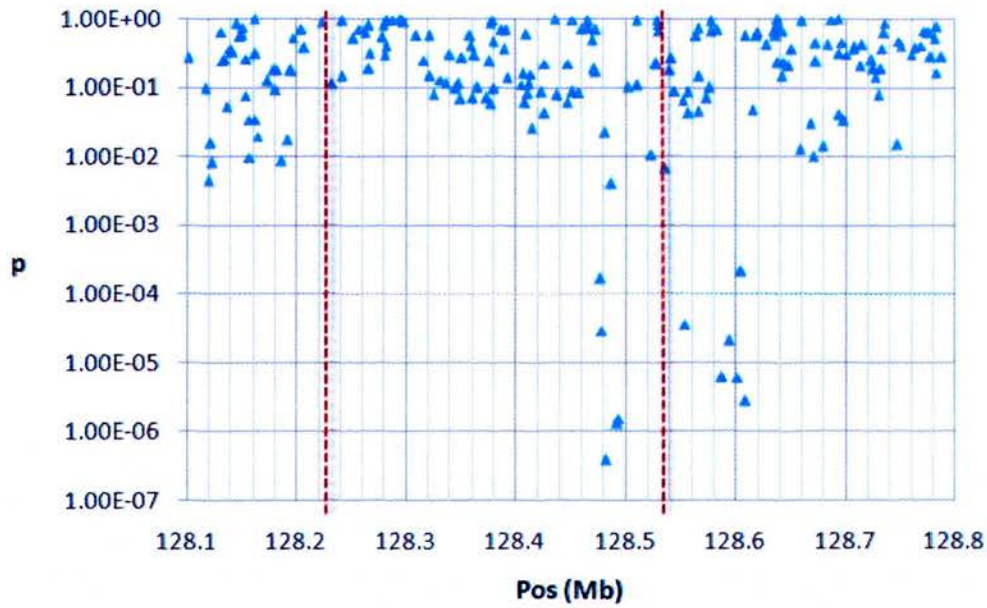


Figure 7.1: Colon, prostate and breast combined p values across the 8q24 locus. The locus is split into the three independent regions as determined by Witte et al.

As can be seen in figure 7.1 the middle of the three independently associated regions determined by Witte et al. appears to show the strongest association with cancer risk when all three tissue types are examined. Table 7.3 illustrates that SNPs in this region are significantly associated with cancer risk in all three individual tissue types with the strongest signal in the prostate study. This is however not the case in the other two regions, with the centromeric region showing no association in prostate tumours and the telomeric region showing no association with breast tumours. It should be noted however that the signal from the colon study is weak in the telomeric region and the signals from both the breast and colon analyses are weak in the centromeric region. Consequently only the central portion of this locus is convincingly associated with all three tumour types. This result is in contradiction to the result of Shumacher et al. [229] who were unable to identify an association between this locus and breast cancer. However this is a result of the fact that they had only typed one polymorphism that was located within the telomeric region of the locus.

Region	rs Id	Pos	Breast p	Prostate p	Colon p	Combined p
centromeric	rs1456306	128185682	0.022	0.27	0.030	0.0085
middle	rs7837328	128492309	0.0037	0.00053	0.0033	$1x10^{-6}$
telomeric	rs11988857	128601055	0.62	$1x10^{-6}$	0.047	$6x10^{-6}$

Table 7.3: Results of representative SNPs from each region

These results consequently argue for a factor within the central portion of this locus that has a common affect on all three cancer types. Unfortunately further analysis suggests that this is not the case. As shown in figure 7.2 those alleles associated with an increased risk in colon and prostate cancers are associated with a decreased risk to breast cancer. Consequently it appears alleles that are protective in breast cancers are deleterious in colon and prostate malignancies (and vice versa). Why this should be the case is unclear and requires further investigation. It may be that the LD structure is different between these three Caucasian populations at this locus. Or alternatively expression patterns of corresponding genes may differ between the tissue types examined.

In order to further narrow down the key region within this locus, we fine-mapped the central portion in our Phase II individuals. To combine the results from our Phase I and II studies, as well as to further characterise the immediate region, we used the IMPUTE program [209] to predict missing genotypes. Only SNPs with a

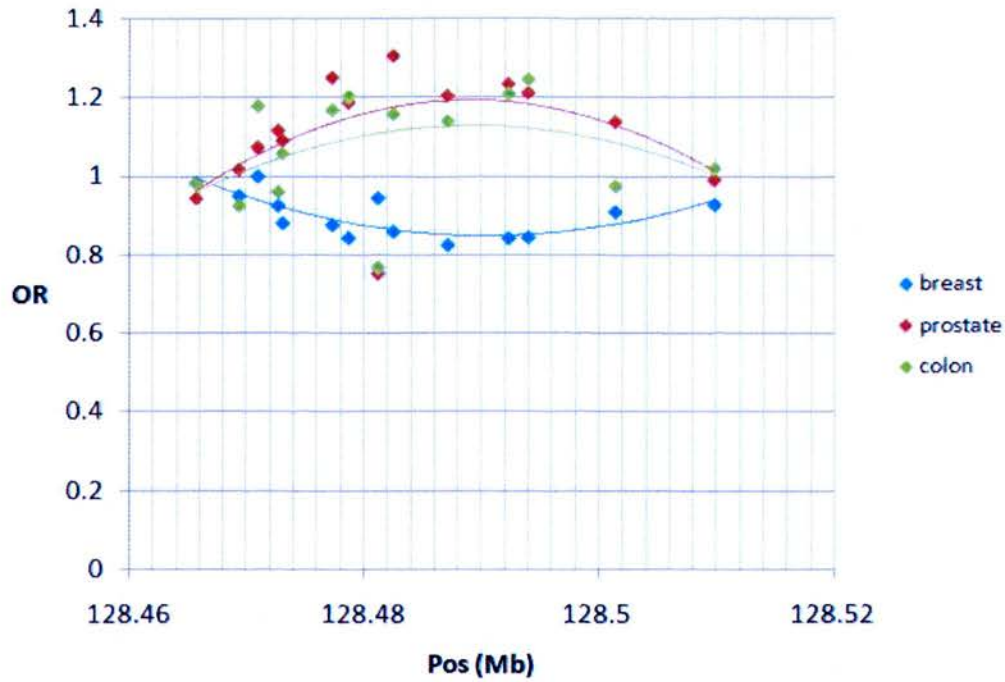


Figure 7.2: The odds ratios, at each SNP, within the middle region of the 8q24 locus. All odds ratios were converted so that the protective allele in breast tumours was examined.

large proportion of high confidence genotypes were retained (>90% p>0.9). To subsequently calculate p values at each SNP the SNPTEST program was used. As can be seen from figure 7.3 the strongest signal from the 8q24 locus is centred within an LD block bounded by two recombination hotspots. This LD block is approximately 63kb in length and contains the *POU5F1L1* gene that is located approximately 2kb centromeric of the polymorphism displaying the strongest signal in our fine-mapping analysis. However a further gene of unknown function, DQ515898, spans this region and there is some EST and transcript evidence that further genes may exist within this LD block. Consequently the exons of these genes are being sequenced by Dr Farrington for new polymorphisms so that an even higher resolution scan can be carried out.

7.3.1 Genome-wide Copy Number Variation Analysis

The use of genotype data to further characterise the mechanisms of disease is not restricted to association studies. Genetic instabilities are a hallmark of the majority of human cancers, and the availability of genome-wide genotype data provides a powerful opportunity to detect chromosomal aberrations such as copy number changes and loss of heterozygosity (LOH). Although classical methods for determining genomic rearrangements and copy number changes, such as FISH (Fluorescent In Situ Hybridisation) and CGH (Comparative Genomic Hybridisation), have had a number of notable successes in characterising genes associated with cancer (e.g. *RB1* [244]), they lack the resolution now possible with genome-wide SNP genotype platforms. We therefore examined the chromosomal aberrations observed in normal individuals and those affected by cancer, in an attempt to determine genomic regions potentially associated with the development of colorectal tumours.

Copy number scores for each SNP were calculated using the Beadstudio Genotyping module from Illumina. The underlying principle of this programs calculation of copy number scores is that regions of abnormal copy number will also display abnormal B allele intensities relative to A allele intensities. For example, in a region with a standard copy number of 2, the B allele intensity relative to the A allele intensity of the SNPs in that region can be thought of as either 0 (no B signal i.e. a AA homozygote), 0.5 (50% B signal i.e. a heterozygote) or 1 (A BB homozygote). However where a copy number of 1 exists only signals of 0 (A) and 1 (B) will be observed. Likewise a copy number of 3 will lead to four potential signals at

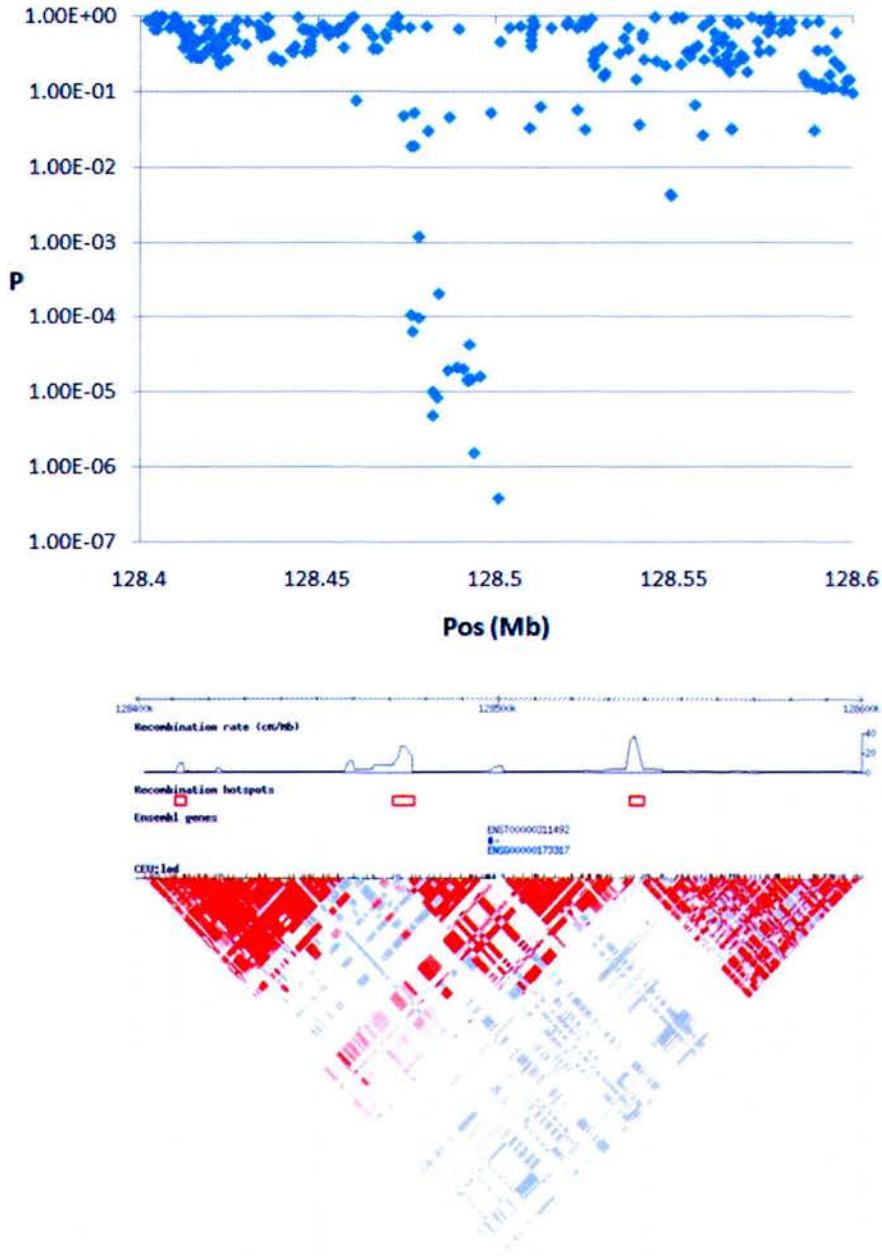


Figure 7.3: 8q24 locus fine-mapping results.
Displayed in the top panel are genotypic p values as calculated by SNPTEST. LD structure and recombination hotspots obtained from HapMap are shown below.

0 (AAA), 0.33 (AAB), 0.66 (ABB) and 1 (BBB). A suitable number of SNPs must however be examined across each region; as if only 1 SNP is examined that displays no B signal it is not possible to determine the copy number in that region (as it could be A, AA, AAA etc). A balance is consequently required between examining enough SNPs across each region to confidently determine that regions copy number, and keeping the region small enough so that small areas of copy number variation can be defined. Given approximately 300,000 SNPs had been typed in this study and the fact that the human genome contains approximately 3 billion base pairs, we had on average one SNP every 10kb. As the average MAF of these SNPs was also around 0.23 we would estimate that approximately 35% of genotypes would be heterozygous. Consequently a minimum sliding window size of 250kb (25 SNPs, 9 heterozygotes) appeared to give us sufficient power to confidently detect copy number changes (though we plan, in the future, to examine more closely the affect on the analysis of adjusting window size; it has been suggested by Illumina that window sizes of 100kb or even 50kb may be sufficient).

To determine regions displaying significantly different levels of copy number variation between normal individuals and those with cancer, we first used the Fishers exact test to determine those SNPs at which an unusual numbers of cases displayed evidence of copy number variation. The most significant polymorphisms in this analysis with a p value of $8x10^{-5}$, that mapped to the small arm of chromosome 19, displayed evidence of copy number variation in only one case compared to 18 controls. Of these 18 controls 17 had a copy number of 1 at this region (Fig 7.4). These results consequently suggest a copy number of one at this locus may provide some form of protection from colorectal cancer.

This region, that spans approximately 350kb, is particularly gene dense and contains 15 Ensembl genes of which a number have previously been associated with cancer. For example this locus contains both a ubiquitin gene as well as the leukaemia associated RNA polymerase II elongation factor. However, perhaps the strongest candidates within this region are the proto-oncogene *JunD* and the NSAID-regulated protein *MIC-1*. *JunD*, that has been shown to have an antiapoptotic role within cells, acts as a modulator of the pathways that links *RAS* to *tp53* (OMIM). The potential role of a proto-oncogene at this locus is supported by the observation of a protective affect of a copy number of one at this region; the fewer copies of a proto-oncogene present within an individuals genome, the lower the chance of one of the copies developing into an oncogene. However this is of course based on the assumption that

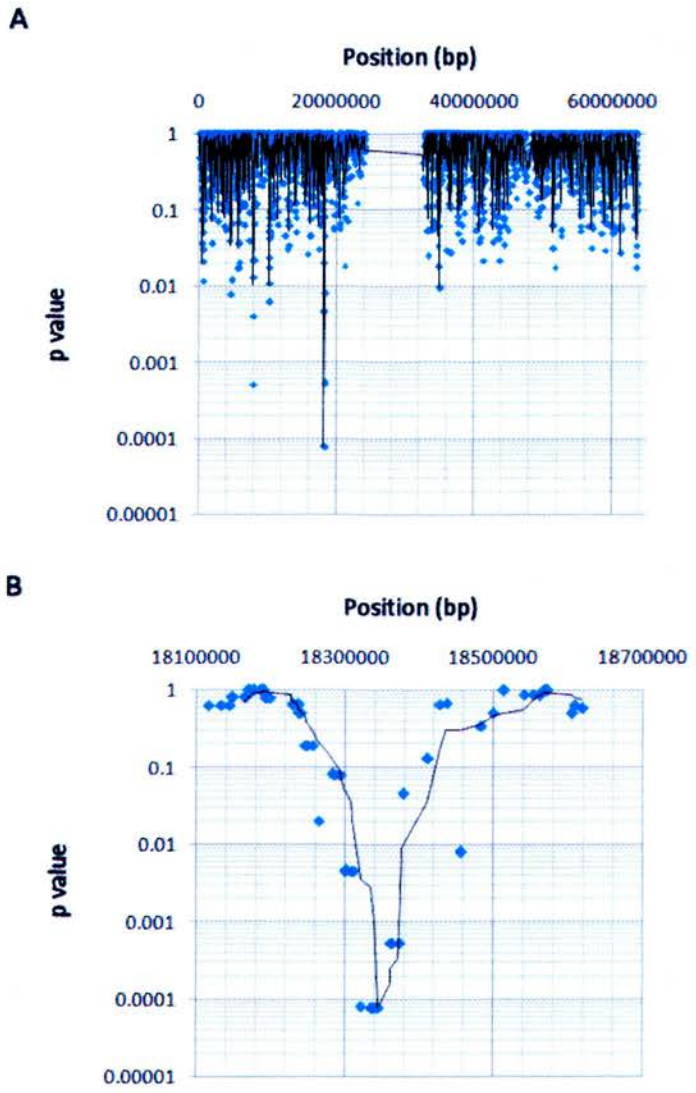


Figure 7.4: Copy number variation across chromosome 19
The copy number variation observed across the whole of chromosome 19 (A) with the most significant region shown in greater detail below (B). The average trendline was calculated using a sliding window size of 5 SNPs.

one copy of the gene is sufficient for life and that any deleterious mutations are dominant. *MIC-1* on the other hand, that is located in the centre of this region, is a member of the transforming growth factor-β (*TGF-*β) superfamily and has generally been associated with antitumorigenic activity and the induction of apoptosis. This would at first appear to contradict the observed enrichment of normal individuals with a copy number of one at this locus, however *MIC-1* has also been shown to exhibit tumourigenic activity under certain conditions and tumour stages (Bauskin et al. [107]). A number of genes in this region were also identified as differentially expressed in tumours through our analysis of ESTs discussed in chapter 2, including ubiquitin and *MIC-1*, and this region is consequently being examined in more detail by Dr Farrington.

Chapter 8

Discussion

In this thesis I aimed to investigate the genetics of colorectal cancer using a variety of approaches. These included the analysis of tumour gene expression across tissues and platforms, the investigation of the association between chromatin structure and cancer gene expression changes, the analysis of mutation and selection rates across differing chromatin conformations as well as the identification and characterisation of cancer risk variants through candidate and whole genome association studies. In summary the main conclusions from this work are:

Chapter 2

- There is limited congruence in tumour gene expression changes observed across different platforms and tissues.
- The human genome contains certain regions that are significantly enriched with genes differentially expressed in tumours.
- The prioritisation of genes for analysis in association studies through the investigation of gene expression changes can be a viable approach (as displayed by the subsequent testing of *EGR1*).

Chapter 3

- Gene expression change in cancers is associated with chromatin structure.
- Genes in open chromatin show the largest increase in expression in cancer. This is observed across platforms and tissues.

- The relationship between chromatin structure and gene expression change appears to be restricted to the later stages of tumourigenesis.
- Genes in more open chromatin display higher levels of general coexpression than those in more closed regions.
- Consequently chromatin appears to be playing an important role in regulating gene expression levels that is dysregulated in cancer.

Chapter 4

- Rates of mutation and selection in the human genome are associated with chromatin structure.
- Background mutation rates are highest in relatively closed regions of the genome, potentially as a result of being less accessible to repair mechanisms.
- Rates of synonymous site divergence are however highest in the most open chromatin regions, a result of being under higher levels of selection in closed regions.
- The more closed the local chromatin structure the greater the difference in the levels of selection experienced by genes with and without CpG islands.
- The local chromatin structure of a gene may, in part, be governed by the rates of mutation and selection it confers.

Chapter 5

- Few, if any, polymorphisms with a previous association to colorectal cancer can be replicated.
- No polymorphisms in DNA repair genes that have a large contribution to colorectal cancer could be identified.
- A weak correlation could be observed between a non-synonymous SNPs association with colorectal cancer and its level of conservation across species.
- Concurrent replication studies can enrich for polymorphisms displaying the strongest evidence of being associated with colorectal cancer, suggesting that

a number of polymorphisms of low risk are likely to be present in DNA repair genes.

Chapter 6

- The GENSCAN program can be successfully used to predict splice variants.
- The only class of polymorphisms that displays significantly higher associations with colorectal cancer than SNPs in general are 5'UTR and regulatory SNPs. Adding further support to the role of gene expression studies in association study gene prioritisation.
- Non-synonymous SNPs in cancer genes do however display a significantly higher distribution of chi-square values than those in genes in general.
- SNP/Gene prioritisation is likely to be limited by the number of true positives among the original dataset.

Chapter 7

- A region on 8q is associated with colorectal, prostate and breast cancer.
- However various regions in this locus appear to modify risk differently in different cancers.
- A copy number variation on chromosome 19 displays evidence of being associated with colorectal cancer.

So what is the relationship between these results and what is their context in the wider field? In this thesis I looked at two examples of prioritisation for association studies; the prioritisation of genes through expression studies and the prioritisation of SNPs by their genomic context. I showed that if different platforms are used to measure expression changes in the same tumour types then there is some level of overlap between the genes deemed differentially expressed. However likewise there is a far greater proportion of genes that are deemed differentially expressed on only one platform. Further work is required to determine whether there is an enrichment of strong candidates among the overlap. With the increase in whole genome association data, this should become increasingly feasible to test.

The subsequent work on *EGR1* does however hold at least some promise for the use of gene expression studies in gene prioritisation. Those genes such as *EGR1* that only have a small affect on cancer incidence (reflected in a small odds ratio) will be hard to detect without some level of pre-prioritisation, as its p value will never survive a multiple test correction if a large number of genes are tested (given a realistic population size). *EGR1* was after all not detected in the whole genome study despite its subsequent replication in further populations examined by Dr Farrington.

On the other hand my examination into SNP prioritisation allowed me to replicate, in a substantially larger dataset, the findings of Zhu et al. that there is a weak association between the test statistic of non-synonymous repair SNPs and their corresponding level of conservation. This is despite none of the SNPs individually reaching significance. This result is therefore promising for those association studies which focus on non-synonymous SNPs, for example the recent publication by the Wellcome Trust Case Control Consortium [230]. However when looking in a genome-wide context, and not at candidate or known cancer associated genes, non-synonymous SNPs as a class did not show substantially higher chi-square values than the majority of SNPs. It appears therefore that typing only non-synonymous SNPs genome-wide is not necessarily a strong strategy, and if only one class of SNPs is to be typed other classes are potentially a better choice, for example 5' and regulatory SNPs. This reflects back on the potential importance of gene expression data in gene prioritisation.

The importance of gene expression in cancer was also reflected in the examination of its relationship to chromatin structure. A number of studies, including that of Zhou et al. [111], have looked in to the distribution of gene expression in cancer and the results suggest chromatin structure may underlie at least part of the non-random distribution found. I proposed that this was a result of common regulatory control between genes in similar chromatin environments, and this hypothesis has been partially supported by other recent studies. For example Stransky et al. [245] looked at the strength of the correlation of each gene in 57 bladder tumours with that of each its neighbouring genes and the genes that show the highest coexpression with their neighbours in this study appear to lie in regions that correspond to areas of open chromatin in the Gilbert et al. analysis [113]. As proposed by Stransky et al. this may reflect common epigenetic alterations that lead to transcriptional deregulation in cancers. Consequently further work is required to investigate this relationship further.

I also proposed that the increased deregulation in cancer in open chromatin regions may be the result of elevated rates of mutation. Loss of efficient DNA repair is a hallmark of many cancer syndromes. However a thorough case control association study of all known DNA repair genes did not produce any strong results despite the use of a relatively large population and replication datasets. This may reflect that although DNA repair loss may underlie many of the rare inherited syndromes in colorectal cancer, the repair pathway may not have such a substantial affect on inherited tumours in the general population.

On the other hand I did show that background mutation rates are likely not constant across the human genome and that DNA repair may be less efficient in closed chromatin. However applying these results directly to cancer will require further work. The mutation spectrum in cancers is still unknown, especially when efficient DNA repair has been lost. This could potentially be addressed by looking at mutations identified through tumour EST libraries and the advent of large scale sequencing technologies such as SOLEXA are likely to assist in investigations of this sort. However we do know that genes in closed chromatin are more likely to undergo mutation in the germ line than those in open chromatin and this consequently has consequences for closed chromatin genes such as *APC*.

Thanks to the advent of the chimpanzee genome I was able to measure these fixed mutation rates by measuring divergence at a number of regions likely to be under little or no selection, including intergenic and intronic regions. However I showed the traditional proxy for background mutation rate, dS, may be unsuitable for this use. By comparing divergence in exons to those in neighbouring introns I showed that divergence at synonymous sites is substantially lower than expected, and that the level of difference is dependent on chromatin structure. Given the clearly defined drop of divergence observed at intron-exon boundaries it is difficult to propose any other explanation for this result than selection occurring at synonymous sites. However this would suggest relatively strong selection is occurring at sites traditionally thought to be under little if any selection, and that this selection is strongest in closed chromatin. Further work is required to try and understand this contradiction.

In the final chapter of this thesis I showed how certain disease loci may be associated with multiple cancers, despite the key regions of these loci differing between tumour types. Likewise I showed that certain changes in a region may be protective in one tumour but deleterious in another. Having mined the results of three association studies pertaining to three different tumour types one locus, found on 8q, was

shown to be associated with all three tumour types. However alleles found to be protective in the breast study were deleterious in colon and prostate tumours.These results illustrate the complex heterogeneity of tumour pathways.

Despite the results listed above this thesis has undoubtedly posed more questions than it has answered. For example what underlies the associations between chromatin structure, tumour gene expression and rates of inter-species divergence? Why do genes in open chromatin display high levels of coexpression and low levels of selection at their synonymous sites? What proportion of disease loci are both protective and deleterious in different tumour types and what implications does this have on selection and SNP prioritisation? Further research is consequently required to expand on the results from this work and to begin to answer some of the questions it has posed.

Chapter 9

Appendix

GO Description	GO ID
-DNA repair	GO:0006281
-base-excision repair	GO:0006284
-base-excision repair\, AP site formation	GO:0006285
-depurination	GO:0045007
-depyrimidination	GO:0045008
-base-excision repair\, base-free sugar-phosphate removal	GO:0006286
-base-excision repair\, DNA ligation	GO:0006288
-base-excision repair\, gap-filling	GO:0006287
-bypass DNA synthesis	GO:0019985
-DNA dealkylation	GO:0006307
-DNA ligation during DNA repair	GO:0051103
-base-excision repair\, DNA ligation	GO:0006288
-DNA replication proofreading	GO:0045004
-DNA synthesis during DNA repair	GO:0000731
-DNA synthesis during double-strand break repair via homologous recombination	GO:0043150
-DNA synthesis during double-strand break repair via single-strand annealing	GO:0043151
-gene conversion at mating-type locus\, DNA repair synthesis	GO:0000734
-meiotic DNA repair synthesis	GO:0000711
-double-strand break repair	GO:0006302
-double-strand break repair via homologous recombination	GO:0000724
-DNA synthesis during double-strand break repair via homologous recombination	GO:0043150

GO Description	GO ID
-double-strand break repair via break-induced replication	GO:0000727
-double-strand break repair via synthesis-dependent strand annealing	GO:0045003
-DNA double-strand break processing	GO:0000729
-gene conversion at mating-type locus\, DNA double-strand break processing	GO:0031292
-meiotic DNA double-strand break processing	GO:0000706
-DNA recombinase assembly	GO:0000730
-meiotic DNA recombinase assembly	GO:0000707
-meiotic recombination nodule assembly	GO:0007146
-early meiotic recombination nodule assembly	GO:0042139
-late meiotic recombination nodule assembly	GO:0042140
-heteroduplex formation	GO:0030491
-meiotic heteroduplex formation	GO:0000713
-strand displacement	GO:0000732
-meiotic strand displacement	GO:0000714
-strand invasion	GO:0042148
-meiotic strand invasion	GO:0000708
-double-strand break repair via nonhomologous end-joining	GO:0006303
-double-strand break repair via single-strand annealing	GO:0045002
-DNA double-strand break processing	GO:0000729
-gene conversion at mating-type locus\, DNA double-strand break processing	GO:0031292
-meiotic DNA double-strand break processing	GO:0000706
-DNA strand renaturation	GO:0000733
-DNA synthesis during double-strand break repair via single-strand annealing	GO:0043151
-double-strand break repair via single-strand annealing\, removal of nonhomologous ends	GO:0000736
-error-free DNA repair	GO:0045021
-error-free postreplication DNA repair	GO:0042275
-error-prone DNA repair	GO:0045020
-error-prone postreplication DNA repair	GO:0042276
-mismatch repair	GO:0006298
-long patch mismatch repair system	GO:0006300
-meiotic mismatch repair	GO:0000710
-short patch mismatch repair system	GO:0006299

GO Description	GO ID
-non-photoreactive DNA repair	GO:0010213
-non-recombinational repair	GO:0000726
-double-strand break repair via nonhomologous end-joining	GO:0006303
-double-strand break repair via single-strand annealing	GO:0045002
-DNA double-strand break processing	GO:0000729
-gene conversion at mating-type locus\, DNA double-strand break processing	GO:0031292
-meiotic DNA double-strand break processing	GO:0000706
-DNA strand renaturation	GO:0000733
-DNA synthesis during double-strand break repair via single-strand annealing	GO:0043151
-double-strand break repair via single-strand annealing\, removal of nonhomologous ends	GO:0000736
-nucleotide-excision repair	GO:0006289
-nucleotide-excision repair\, DNA damage recognition	GO:0000715
-transcription-coupled nucleotide-excision repair\, DNA damage recognition	GO:0000716
-nucleotide-excision repair\, DNA damage removal	GO:0000718
-nucleotide-excision repair\, DNA duplex unwinding	GO:0000717
-nucleotide-excision repair\, DNA gap filling	GO:0006297
-nucleotide-excision repair\, DNA incision\, 3'-to lesion	GO:0006295
-nucleotide-excision repair\, DNA incision\, 5'-to lesion	GO:0006296
-nucleotide-excision repair\, preincision complex formation	GO:0006294
-nucleotide-excision repair\, preincision complex stabilization	GO:0006293
-pyrimidine dimer repair via nucleotide excision repair	GO:0000720
-transcription-coupled nucleotide-excision repair	GO:0006283
-transcription-coupled nucleotide-excision repair\, DNA damage recognition	GO:0000716
-postreplication repair	GO:0006301
-error-free postreplication DNA repair	GO:0042275
-error-prone postreplication DNA repair	GO:0042276
-pyrimidine dimer repair	GO:0006290
-photoreactive repair	GO:0000719
-pyrimidine dimer repair via nucleotide excision repair	GO:0000720
-recombinational repair	GO:0000725
-double-strand break repair via homologous recombination	GO:0000724
-DNA synthesis during double-strand break repair via homologous recombination	GO:0043150

GO Description	GO ID
-double-strand break repair via break-induced replication	GO:0000727
-double-strand break repair via synthesis-dependent strand annealing	GO:0045003
-DNA double-strand break processing	GO:0000729
-gene conversion at mating-type locus\, DNA double-strand break processing	GO:0031292
-meiotic DNA double-strand break processing	GO:0000706
-DNA recombinase assembly	GO:0000730
-meiotic DNA recombinase assembly	GO:0000707
-meiotic recombination nodule assembly	GO:0007146
-early meiotic recombination nodule assembly	GO:0042139
-late meiotic recombination nodule assembly	GO:0042140
-heteroduplex formation	GO:0030491
-meiotic heteroduplex formation	GO:0000713
-strand displacement	GO:0000732
-meiotic strand displacement	GO:0000714
-strand invasion	GO:0042148
-meiotic strand invasion	GO:0000708
-regulation of DNA repair	GO:0006282
-negative regulation of DNA repair	GO:0045738
-positive regulation of DNA repair	GO:0045739
-single strand break repair	GO:0000012
-viral DNA repair	GO:0046787
-recruitment of helicase-primase complex to DNA lesions	GO:0046799

Table 9.1: DNA repair associated GO terms

Ensembl ID	HGNC	GO description
ENSG00000097007	ABL1	mismatch repair
ENSG00000100601	ALKBH	DNA repair // DNA dealkylation
ENSG00000132466	ANKRD17	mismatch repair
ENSG00000100823	APEX1	base-excision repair
ENSG00000169188	APEX2	DNA repair
ENSG00000137074	APTX	single strand break repair // base-excision repair
ENSG00000034533	ASTE1	DNA repair

Ensembl ID	HGNC	GO description
ENSG00000149311	ATM	DNA repair
ENSG00000175054	ATR	DNA repair
ENSG00000085224	ATRX	DNA repair
ENSG00000197299	BLM	DNA repair
ENSG00000012048	BRCA1	DNA repair // positive regulation of DNA repair
ENSG00000139618	BRCA2	double-strand break repair via homologous recombination // DNA repair
ENSG00000136492	BRIP1	double-strand break repair
ENSG00000159388	BTG2	DNA repair
ENSG00000134480	CCNH	DNA repair
ENSG00000134058	CDK7	DNA repair
ENSG00000147400	CETN2	
ENSG00000167670	CHAF1A	DNA repair
ENSG00000159259	CHAF1B	DNA repair
ENSG00000149554	CHEK1	DNA repair
ENSG00000183765	CHEK2	
ENSG00000185043	CIB1	double-strand break repair
ENSG00000008405	CRY1	DNA repair
ENSG00000121671	CRY2	DNA repair
ENSG00000141551	CSNK1D	DNA repair
ENSG00000100181	CSNK1E	DNA repair
ENSG00000108055	CSPG6	DNA repair
ENSG00000198924	DCLRE1A	DNA repair // nucleotide-excision repair
ENSG00000118655	DCLRE1B	DNA repair
ENSG00000152457	DCLRE1C	DNA repair
ENSG00000167986	DDB1	nucleotide-excision repair
ENSG00000134574	DDB2	nucleotide-excision repair // pyrimidine dimer repair
ENSG00000013573	DDX11	
ENSG00000111788	DDX12	
ENSG00000100206	DMC1	
ENSG00000130816	DNMT1	
ENSG00000107614	DNMT2	

Ensembl ID	HGNC	GO description
ENSG00000119772	DNMT3A	
ENSG00000088305	DNMT3B	
ENSG00000142182	DNMT3L	
ENSG00000128951	DUT	
ENSG00000154920	EME1	
ENSG00000012061	ERCC1	DNA repair // nucleotide-excision repair
ENSG00000104884	ERCC2	transcription-coupled nucleotide-excision repair
ENSG00000163161	ERCC3	transcription-coupled nucleotide-excision repair
ENSG00000175595	ERCC4	nucleotide-excision repair
ENSG00000134899	ERCC5	DNA repair // transcription-coupled nucleotide-excision repair // nucleotide-excision repair
ENSG00000049167	ERCC8	DNA repair // transcription-coupled nucleotide-excision repair
ENSG00000174371	EXO1	DNA repair // mismatch repair
ENSG00000187741	FANCA	DNA repair
ENSG00000181544	FANCB	DNA repair
ENSG00000158169	FANCC	DNA repair // nucleotide-excision repair
ENSG00000144554	FANCD2	DNA repair
ENSG00000112039	FANCE	DNA repair
ENSG00000183161	FANCF	DNA repair
ENSG00000165281	FANCG	DNA repair
ENSG00000115392	FANCL	DNA repair
ENSG00000134452	FBXO18	DNA repair
ENSG00000168496	FEN1	DNA repair // double-strand break repair
ENSG00000198793	FRAP1	
ENSG00000116717	GADD45A	DNA repair
ENSG00000130222	GADD45G	DNA repair
ENSG00000110768	GTF2H1	DNA repair
ENSG00000145736	GTF2H2	DNA repair
ENSG00000111358	GTF2H3	nucleotide-excision repair
ENSG00000137349	GTF2H4	DNA repair
ENSG00000185068	GTF2H5	DNA repair

Ensembl ID	HGNC	GO description
ENSG00000188486	H2AFX	double-strand break repair via homologous recombination // DNA repair
ENSG00000100118	HMG1L10	DNA repair // base-excision repair, DNA ligation
ENSG00000189403	HMGB1	DNA repair // base-excision repair, DNA ligation
ENSG00000164104	HMGB2	DNA repair // base-excision repair, DNA ligation
ENSG00000172977	HTATIP	double-strand break repair
ENSG00000136273	HUS1	DNA repair
ENSG00000178922	HYI	
ENSG00000132740	IGHMBP2	DNA repair
ENSG00000161896	IHPK3	DNA repair
ENSG00000198690	KIAA1018	DNA repair
ENSG00000105486	LIG1	DNA repair
ENSG00000005156	LIG3	DNA repair
ENSG00000174405	LIG4	single strand break repair // DNA repair
ENSG00000116670	MAD2L2	
ENSG00000129071	MBD4	base-excision repair
ENSG00000170430	MGMT	DNA repair // DNA dealkylation
ENSG00000076242	MLH1	mismatch repair
ENSG00000119684	MLH3	mismatch repair
ENSG00000155229	MMS19L	DNA repair
ENSG00000020426	MNAT1	DNA repair
ENSG00000103152	MPG	base-excision repair // DNA dealkylation
ENSG00000020922	MRE11A	double-strand break repair via nonhomologous end-joining
ENSG00000095002	MSH2	mismatch repair // postreplication repair
ENSG00000113318	MSH3	mismatch repair
ENSG00000057468	MSH4	mismatch repair
ENSG00000096474	MSH5	mismatch repair
ENSG00000116062	MSH6	mismatch repair // short patch mismatch repair system
ENSG00000172732	MUS81	DNA repair
ENSG00000132781	MUTYH	base-excision repair // mismatch repair
ENSG00000196535	MYO18A	
ENSG00000104320	NBN	double-strand break repair

Ensembl ID	HGNC	GO description
ENSG00000198646	NCOA6	DNA repair
ENSG00000140398	NEIL1	DNA repair
ENSG00000154328	NEIL2	DNA repair
ENSG00000109674	NEIL3	DNA repair
ENSG00000065057	NTHL1	base-excision repair // nucleotide-excision repair, DNA incision, 5'-to lesion
ENSG00000106268	NUDT1	DNA repair
ENSG00000114026	OGG1	base-excision repair
ENSG00000143799	PARP1	DNA repair // base-excision repair
ENSG00000129484	PARP2	DNA repair // base-excision repair
ENSG00000041880	PARP3	DNA repair
ENSG00000102699	PARP4	DNA repair
ENSG00000132646	PCNA	DNA repair // base-excision repair, gap-filling
ENSG00000032514	PGBD3	DNA repair // transcription-coupled nucleotide-excision repair // pyrimidine dimer repair
ENSG00000064933	PMS1	mismatch repair
ENSG00000122512	PMS2	mismatch repair
ENSG00000078319	PMS2L1	mismatch repair
ENSG00000186704	PMS2L11	mismatch repair
ENSG00000127957	PMS2L3	mismatch repair
ENSG00000067601	PMS2L4	mismatch repair
ENSG00000039650	PNKP	nucleotide-excision repair, DNA damage removal // DNA repair
ENSG00000070501	POLB	DNA repair // base-excision repair, gap-filling
ENSG00000062822	POLD1	DNA repair // base-excision repair, gap-filling // DNA replication proofreading
ENSG00000077514	POLD3	DNA synthesis during DNA repair // mismatch repair
ENSG00000177084	POLE	DNA repair
ENSG00000100479	POLE2	DNA repair
ENSG00000140521	POLG	base-excision repair, gap-filling
ENSG00000136480	POLG2	DNA repair
ENSG00000170734	POLH	DNA repair // regulation of DNA repair // pyrimidine dimer repair // postreplication repair

Ensembl ID	HGNC	GO description
ENSG00000101751	POLI	DNA repair
ENSG00000122008	POLK	DNA repair
ENSG00000166169	POLL	DNA repair // nucleotide-excision repair
ENSG00000122678	POLM	
ENSG00000130997	POLN	
ENSG00000051341	POLQ	DNA repair
ENSG00000121031	PRKDC	double-strand break repair
ENSG00000110107	PRPF19	DNA repair
ENSG00000164611	PTTG1	DNA repair
ENSG00000113456	RAD1	DNA repair
ENSG00000152942	RAD17	DNA repair
ENSG00000070950	RAD18	DNA repair
ENSG00000164754	RAD21	double-strand break repair
ENSG00000179262	RAD23A	nucleotide-excision repair
ENSG00000119318	RAD23B	nucleotide-excision repair
ENSG00000113522	RAD50	double-strand break repair
ENSG00000051180	RAD51	double-strand break repair via homologous recombination // DNA repair
ENSG00000111247	RAD51AP1	double-strand break repair via homologous recombination // DNA repair
ENSG00000108384	RAD51C	DNA repair
ENSG00000182185	RAD51L1	DNA repair
ENSG00000185379	RAD51L3	base-excision repair
ENSG00000002016	RAD52	DNA repair // double-strand break repair via homologous recombination // double-strand break repair
ENSG00000197275	RAD54B	DNA repair
ENSG00000085999	RAD54L	DNA repair
ENSG00000172613	RAD9A	DNA repair
ENSG00000151164	RAD9B	DNA repair
ENSG00000162521	RBBP4	DNA repair
ENSG00000101773	RBBP8	DNA repair
ENSG00000173959	RBM14	DNA repair

Ensembl ID	HGNC	GO description
ENSG00000187456	RDM1	
ENSG00000004700	RECQL	DNA repair
ENSG00000160957	RECQL4	DNA repair
ENSG00000108469	RECQL5	DNA repair
ENSG00000135945	REV1L	DNA repair // error-prone postreplication DNA repair
ENSG00000009413	REV3L	DNA repair
ENSG00000111445	RFC5	DNA repair
ENSG00000132383	RPA1	DNA repair
ENSG00000117748	RPA2	
ENSG00000106399	RPA3	DNA repair
OTTHUMG00000021987	RPA4	
ENSG00000048392	RRM2B	
ENSG00000183207	RUVBL2	DNA repair
ENSG00000127922	SHFM1	
ENSG00000065613	SLK	nucleotide-excision repair
ENSG00000072501	SMC1L1	DNA repair
ENSG00000054796	SPO11	
ENSG00000132207	SULT1A3	DNA repair
ENSG00000181625	SULT1A3	DNA repair
ENSG00000139372	TDG	DNA repair // base-excision repair
ENSG00000042088	TDP1	DNA repair
ENSG00000026036	TNFRSF6B	
ENSG00000118245	TNP1	single strand break repair
ENSG00000141510	TP53	base-excision repair // nucleotide-excision repair
ENSG00000078900	TP73	mismatch repair
ENSG00000164053	TREX1	DNA repair // mismatch repair
ENSG00000183479	TREX2	DNA repair
ENSG00000077721	UBE2A	postreplication repair
ENSG00000119048	UBE2B	DNA repair
ENSG00000177889	UBE2N	
ENSG00000169139	UBE2V2	regulation of DNA repair
ENSG00000076248	UNG	DNA repair // base-excision repair

Ensembl ID	HGNC	GO description
ENSG00000152669	UNG2	base-excision repair
ENSG00000136709	WDR33	postreplication repair
ENSG00000165392	WRN	
ENSG00000124535	WRNIP1	DNA synthesis during DNA repair
ENSG00000076924	XAB2	DNA repair // transcription-coupled nucleotide-excision repair
ENSG00000136936	XPA	nucleotide-excision repair
ENSG00000154767	XPC	nucleotide-excision repair
ENSG00000073050	XRCC1	single strand break repair
ENSG00000196584	XRCC2	DNA repair
ENSG00000126215	XRCC3	DNA repair
ENSG00000152422	XRCC4	DNA repair // double-strand break repair
ENSG00000079246	XRCC5	regulation of DNA repair // double-strand break repair via nonhomologous end-joining
ENSG00000196419	XRCC6	DNA repair // double-strand break repair via nonhomologous end-joining
ENSG00000088930	XRN2	DNA repair
ENSG00000010072		DNA repair
ENSG00000078177		mismatch repair
ENSG00000123415		DNA repair
ENSG00000123965		mismatch repair
ENSG00000131944		
ENSG00000135165		mismatch repair
ENSG00000158636		DNA repair
ENSG00000163312		
ENSG00000166199		
ENSG00000166896		double-strand break repair via nonhomologous end-joining
ENSG00000173818		DNA repair
ENSG00000174368		mismatch repair
ENSG00000178295		DNA repair
ENSG00000187953		mismatch repair
ENSG00000189046		

Ensembl ID	HGNC	GO description
ENSG00000196120		
ENSG00000196763		single strand break repair
ENSG00000197229		base-excision repair // nucleotide-excision repair

Table 9.2: DNA repair genes

SNP	Gene	Reference
rs1042522	TP53	Dumont P, 2003[246]
rs1045642	MDR1	Jamroziak K, 2004[247]
rs1047840	EXO1	Wu Y, 2001[248]
rs1047972	AURKA	Kemp Z, 2004[41]
rs1048943	CYP1A1	Kiyohara C, 2000[249]
rs1049654	CD36	Kemp Z, 2004[41]
rs1051740	mEPHX	Harrison DJ 1999[250]
rs1056827	CYP1B1	Kiyohara C, 2000[249]
rs1056836	CYP1B1	Kiyohara C, 2000[249]
rs1057910	CYP2C9	Kemp Z, 2004[41]
rs1059060	PMS2	-
rs10735810	VDR	Kemp Z, 2004[41]
rs10916	CYP1B1	Kiyohara C, 2000[249]
rs1128503	MDR1	-
rs1143627	IL1B	El Omar, 2000[251]
rs11692021	UGT1A7	Strassburg CP, 2002[252]
rs12917	MGMT	Kemp Z, 2004[41]
rs13181	XPD	Yeh CC, 2005[253]
rs1503185	PTPRJ	Kemp Z, 2004[41]
rs1566734	PTPRJ	Kemp Z, 2004[41]
rs16260	CDH1	Porter TR, 2002[254]
rs1650697	MSH3	Orimo H 2000[255]
rs16944	IL1B	Kemp Z, 2004[41]
rs175080	MLH3	de Jong MM 2004[256]
rs17561	IL1A	Yoshimura K, 2003[257]
rs17868323	UGT1A7	Strassburg CP, 2002[252]

SNP	Gene	Reference
rs17879961	CHEK2	Kemp Z, 2004[41]
rs1799782	XRCC1	Shen H, 2000[258]
rs1799930	NAT2	Brockton N, 2000[259]
rs1799931	NAT2	Brockton N, 2000[259]
rs1799945	HFE	Kemp Z, 2004[41]
rs1799977	MLH1	Bagnoli S, 2004[260]
rs1800562	HFE	Kemp Z, 2004[41]
rs1800566	NQO1	Kemp Z, 2004[41]
rs1800629	TNF	-
rs1800795	IL6	Kemp Z, 2004[41]
rs1801131	MTHFR	Kiyohara C, 2000[249]
rs1801133	MTHFR	Kiyohara C, 2000[249]
rs1801278	IRS1	Slattery ML, 2004[261]
rs1801280	NAT2	Brockton N, 2000[259]
rs1801282	PPARG	Landi S, 2003[262]
rs1801376	BUB1B	Cahill DP, 1998[263]
rs1801394	FASTKD3	Kemp Z, 2004[41]
rs1805321	PMS2	-
rs1865434	IRS2	Slattery ML, 2004[261]
rs20417	PTGS2	Kemp Z, 2004[41]
rs2066844	NOD2	Hampe, 2002[264]
rs2066845	NOD2	Hampe, 2002[264]
rs2234922	mEPHX	Harrison DJ 1999[250]
rs2273535	STK15	Ewart-Toland A, 2003[265]
rs2276331	CDH1	Kim HC, 2000[266]
rs2287498	TP53	Dumont P, 2003[246]
rs2303428	MSH2	Kolodner RD, 1994[267]
rs25487	XRCC1	Krupa R, 2004[268]
rs25489	XRCC1	Kemp Z, 2004[41]
rs2854744	IGFBP3	Slattery ML, 2004[261]
rs361525	TNF	-
rs3750861	KLF6	Kemp Z, 2004[41]

SNP	Gene	Reference
rs3808607	CYP7A1	Kemp Z, 2004[41]
rs4073	IL8	Kemp Z, 2004[41]
rs4148323	UGT1A1	Kemp Z, 2004[41]
rs4149963	EXO1	Wu Y, 2001[248]
rs4149966	EXO1	Wu Y, 2001[248]
rs429358	APOE	Kemp Z, 2004[41]
rs4648298	COX2	Cox DG, 2004[269]
rs4988235	LCT	-
rs649392	CCND1	Kong S, 2000[270]
rs689469	COX2	Cox DG, 2004[269]
rs735943	EXO1	Wu Y, 2001[248]
rs7412	APOE	Kemp Z, 2004[41]
rs7586110	UGT1A1	Kemp Z, 2004[41]
rs861539	XRCC3	Krupa R, 2004[268]
rs9350	EXO1	Wu Y, 2001[248]

Table 9.3: Polymorphisms with a previous link to colorectal cancer

Gene
EEF1A2
IGF1
SDHALP2
GSTP1
ABCB1
FH
Rap2-binding protein 9
TGFBR1
PPARD
SDHB
PTPRJ
NAT2
CBS

Gene
HIF1AN
AXIN2
TYMS
WAF-1/CIP1 stabilizing protein 39
LGALS1
LGALS10
LGALS12
LGALS13
LGALS14
LGALS2
LGALS3
LGALS4
LGALS7
LGALS8
LGALS9

Table 9.4: Non-repair candidate genes examined

Bibliography

[1] Statistics ISC: **http://www.isdscotland.org/isd/1425.html**.

[2] incidence statistics C: **http://info.cancerresearchuk.org/cancerstats/types/bowel/incidence/?a=5441**.

[3] stage statistics C: **http://info.cancerresearchuk.org/cancerstats/types/bowel/screeningandprevention/?a=5441**.

[4] Hardy RG, Meltzer SJ, Jankowski JA: **ABC of colorectal cancer. Molecular basis for risk factors.** *BMJ* 2000, **321**(7265):886–889.

[5] Vogelstein B, Kinzler KW: **The multistep nature of cancer.** *Trends Genet* 1993, **9**(4):138–141.

[6] Kelloff GJ, Schilsky RL, Alberts DS, Day RW, Guyton KZ, Pearce HL, Peck JC, Phillips R, Sigman CC: **Colorectal adenomas: a prototype for the use of surrogate end points in the development of cancer prevention drugs.** *Clin Cancer Res* 2004, **10**(11):3908–3918.

[7] Jass JR, Whitehall VLJ, Young J, Leggett BA: **Emerging concepts in colorectal neoplasia.** *Gastroenterology* 2002, **123**(3):862–876.

[8] Clevers H: **At the crossroads of inflammation and cancer.** *Cell* 2004, **118**(6):671–674.

[9] Chan TL, Curtis LC, Leung SY, Farrington SM, Ho JW, Chan AS, Lam PW, Tse CW, Dunlop MG, Wyllie AH, Yuen ST: **Early-onset colorectal cancer with stable microsatellite DNA and near-diploid chromosomes.** *Oncogene* 2001, **20**(35):4871–4876.

[10] Chung DC, Rustgi AK: **The hereditary nonpolyposis colorectal cancer syndrome: genetics and clinical implications.** *Ann Intern Med* 2003, **138**(7):560–570.

[11] Jacob S, Praz F: **DNA mismatch repair defects: role in colorectal carcinogenesis.** *Biochimie* 2002, **84**:27–47.

[12] Jiricny J, Marra G: **DNA repair defects in colon cancer.** *Curr Opin Genet Dev* 2003, **13**:61–69.

[13] de la Chapelle A: **Microsatellite instability.** *N Engl J Med* 2003, **349**(3):209–210.

[14] Grady WM: **Genomic instability and colon cancer.** *Cancer Metastasis Rev* 2004, **23**(1-2):11–27.

[15] Lengauer C, Kinzler KW, Vogelstein B: **Genetic instabilities in human cancers.** *Nature* 1998, **396**(6712):643–649.

[16] John DJS, McDermott FT, Hopper JL, Debney EA, Johnson WR, Hughes ES: **Cancer risk in relatives of patients with common colorectal cancer.** *Ann Intern Med* 1993, **118**(10):785–790.

[17] Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC: **A prospective study of family history and the risk of colorectal cancer.** *N Engl J Med* 1994, **331**(25):1669–1674.

[18] Slattery ML, Kerber RA: **Family history of cancer and colon cancer risk: the Utah Population Database.** *J Natl Cancer Inst* 1994, **86**(21):1618–1626.

[19] Bussey H: *Familial Polyposis Coli: Family Studies, Histopathology, Differential Diagnosis and Results of Treatment.* The Johns Hopkins University Press 1975.

[20] Hyer W, Fell JM: **Screening for familial adenomatous polyposis.** *Arch Dis Child* 2001, **84**(5):377–380.

[21] Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR, Cheadle JP: **Inherited**

variants of MYH associated with somatic G:C–>T:A mutations in colorectal tumors. *Nat Genet* 2002, **30**(2):227–232.

[22] Sieber OM, Lipton L, Crabtree M, Heinimann K, Fidalgo P, Phillips RKS, Bisgaard ML, Orntoft TF, Aaltonen LA, Hodgson SV, Thomas HJW, Tomlinson IPM: **Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH.** *N Engl J Med* 2003, **348**(9):791–799.

[23] Sampson JR, Dolwani S, Jones S, Eccles D, Ellis A, Evans DG, Frayling I, Jordan S, Maher ER, Mak T, Maynard J, Pigatto F, Shaw J, Cheadle JP: **Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH.** *Lancet* 2003, **362**(9377):39–41.

[24] Galiatsatos P, Foulkes WD: **Familial adenomatous polyposis.** *Am J Gastroenterol* 2006, **101**(2):385–398.

[25] Vasen HF, Wijnen JT, Menko FH, Kleibeuker JH, Taal BG, Griffioen G, Nagengast FM, Meijers-Heijboer EH, Bertario L, Varesco L, Bisgaard ML, Mohr J, Fodde R, Khan PM: **Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis.** *Gastroenterology* 1996, **110**(4):1020–1027.

[26] Söreide K, Janssen EAM, Söiland H, Körner H, Baak JPA: **Microsatellite instability in colorectal cancer.** *Br J Surg* 2006, **93**(4):395–406.

[27] Lynch HT, de la Chapelle A: **Genetic susceptibility to non-polyposis colorectal cancer.** *J Med Genet* 1999, **36**(11):801–818.

[28] Jeghers H, McKusick VA, Katz KH: **Generalized intestinal polyposis and melanin spots of the oral mucosa, lips and digits; a syndrome of diagnostic significance.** *N Engl J Med* 1949, **241**(26):1031–1036.

[29] Hearle N, Schumacher V, Menko FH, Olschwang S, Boardman LA, Gille JJP, Keller JJ, Westerman AM, Scott RJ, Lim W, Trimbath JD, Giardiello FM, Gruber SB, Offerhaus GJA, de Rooij FWM, Wilson JHP, Hansmann A, Möslein G, Royer-Pokora B, Vogel T, Phillips RKS, Spigelman AD, Houlston RS: **Frequency and spectrum of cancers in the Peutz-Jeghers syndrome.** *Clin Cancer Res* 2006, **12**(10):3209–3215.

[30] Aretz S, Stienen D, Uhlhaas S, Loff S, Back W, Pagenstecher C, McLeod DR, Graham GE, Mangold E, Santer R, Propping P, Friedl W: **High proportion of large genomic STK11 deletions in Peutz-Jeghers syndrome.** *Hum Mutat* 2005, **26**(6):513–519.

[31] Howe JR, Roth S, Ringold JC, Summers RW, Järvinen HJ, Sistonen P, Tomlinson IP, Houlston RS, Bevan S, Mitros FA, Stone EM, Aaltonen LA: **Mutations in the SMAD4/DPC4 gene in juvenile polyposis.** *Science* 1998, **280**(5366):1086–1088.

[32] Howe JR, Bair JL, Sayed MG, Anderson ME, Mitros FA, Petersen GM, Velculescu VE, Traverso G, Vogelstein B: **Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis.** *Nat Genet* 2001, **28**(2):184–187.

[33] Herrera L, Kakati S, Gibas L, Pietrzak E, Sandberg AA: **Gardner syndrome in a man with an interstitial deletion of 5q.** *Am J Med Genet* 1986, **25**(3):473–476.

[34] Gryfe R, Nicola ND, Gallinger S, Redston M: **Somatic instability of the APC I1307K allele in colorectal neoplasia.** *Cancer Res* 1998, **58**(18):4040–4043.

[35] Bright-Thomas RM, Hargest R: **APC, beta-Catenin and hTCF-4; an unholy trinity in the genesis of colorectal cancer.** *Eur J Surg Oncol* 2003, **29**(2):107–117.

[36] Calvert PM, Frucht H: **The genetics of colorectal cancer.** *Ann Intern Med* 2002, **137**(7):603–612.

[37] Karim R, Tse G, Putti T, Scolyer R, Lee S: **The significance of the Wnt pathway in the pathology of human cancers.** *Pathology* 2004, **36**(2):120–128.

[38] Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL: **Genetic alterations during colorectal-tumor development.** *N Engl J Med* 1988, **319**(9):525–532.

[39] Friedl W, Kruse R, Uhlhaas S, Stolte M, Schartmann B, Keller KM, Jungck M, Stern M, Loff S, Back W, Propping P, Jenne DE: **Frequent 4-bp deletion in exon 9 of the SMAD4/MADH4 gene in familial juvenile polyposis patients.** *Genes Chromosomes Cancer* 1999, **25**(4):403–406.

[40] Lynch HT, de la Chapelle A: **Hereditary colorectal cancer.** *N Engl J Med* 2003, **348**(10):919–932.

[41] Kemp Z, Thirlwell C, Sieber O, Silver A, Tomlinson I: **An update on the genetics of colorectal cancer.** *Hum Mol Genet* 2004, **13 Spec No 2**:R177–R185.

[42] Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland.** *N Engl J Med* 2000, **343**(2):78–85.

[43] de Jong MM, Nolte IM, te Meerman GJ, van der Graaf WTA, de Vries EGE, Sijmons RH, Hofstra RMW, Kleibeuker JH: **Low-penetrance genes and their involvement in colorectal cancer susceptibility.** *Cancer Epidemiol Biomarkers Prev* 2002, **11**(11):1332–1352.

[44] Houlston RS, Peto J: **The search for low-penetrance cancer susceptibility alleles.** *Oncogene* 2004, **23**(38):6471–6476.

[45] Segnan N, Senore C, Andreoni B, Arrigoni A, Bisanti L, Cardelli A, Castiglione G, Crosta C, DiPlacido R, Ferrari A, Ferraris R, Ferrero F, Fracchia M, Gasperoni S, Malfitana G, Recchia S, Risio M, Rizzetto M, Saracco G, Spandre M, Turco D, Turco P, Zappa M, Group-Italy SCOREW: **Randomized trial of different screening strategies for colorectal cancer: patient response and detection rates.** *J Natl Cancer Inst* 2005, **97**(5):347–357.

[46] Stemmermann GN, Mandel M, Mower HF: **Colon cancer: its precursors and companions in Hawaii Japanese.** *Natl Cancer Inst Monogr* 1979, (53):175–179.

[47] Chan AT, Giovannucci EL, Meyerhardt JA, Schernhammer ES, Curhan GC, Fuchs CS: **Long-term use of aspirin and nonsteroidal anti-**

inflammatory drugs and risk of colorectal cancer. *JAMA* 2005, **294**(8):914–923.

[48] Rostom A, Dubé C, Lewin G, Tsertsvadze A, Barrowman N, Code C, Sampson M, Moher D, Force UPST: **Nonsteroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors for primary prevention of colorectal cancer: a systematic review prepared for the U.S. Preventive Services Task Force.** *Ann Intern Med* 2007, **146**(5):376–389.

[49] Nanda K, Bastian LA, Hasselblad V, Simel DL: **Hormone replacement therapy and the risk of colorectal cancer: a meta-analysis.** *Obstet Gynecol* 1999, **93**(5 Pt 2):880–888.

[50] Terry PD, Miller AB, Rohan TE: **Obesity and colorectal cancer risk in women.** *Gut* 2002, **51**(2):191–194.

[51] Kufe DW, Bast RC, Hait W, Hong WK, Pollock RE, Weichselbaum RR, Holland JF, Frei E: *Cancer Medicine.* BC Decker 2006.

[52] Reynolds MA, Kastury K, Groskopf J, Schalken JA, Rittenhouse H: **Molecular markers for prostate cancer.** *Cancer Lett* 2007, **249**:5–13.

[53] Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SAL, Saha D, Goldman L, Petricoin EF, Conrads TP, Veenstra TD, Loffredo CA, Goldman R: **Analysis of mass spectral serum profiles for biomarker selection.** *Bioinformatics* 2005, **21**(21):4039–4045.

[54] Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information.** *Anal Chem* 2000, **72**(11):2482–2489.

[55] Ressom HW, Varghese RS, Drake SK, Hortin GL, Abdel-Hamid M, Loffredo CA, Goldman R: **Peak selection from MALDI-TOF mass spectra using ant colony optimization.** *Bioinformatics* 2007, **23**(5):619–626.

[56] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd

JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**(6769):503–511.

[57] Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**(2):263–265.

[58] Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics* 2002, **18 Suppl 2**:S110–S115.

[59] Perez-Iratxeta C, Wjst M, Bork P, Andrade MA: **G2D: a tool for mining genes associated with disease.** *BMC Genet* 2005, **6**:45.

[60] Turner BM: *Chromatin and gene regulation. Molecular mechanisms in epigenetics.* Blackwell science 2001.

[61] Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005, **6**:55.

[62] Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates.** *Nucleic Acids Res* 2005, **33**(5):1544–1552.

[63] López-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease.** *Nucleic Acids Res* 2004, **32**(10):3108–3114.

[64] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**(5675):1321–1325.

[65] Ban M, Maranian M, Yeo TW, Gray J, Compston A, Sawcer S: **Ultraconserved regions in multiple sclerosis.** *Eur J Hum Genet* 2005, **13**(9):998–999.

[66] Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucl.Acids Res.* 2003, **31**(13):3812–3814.

[67] Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT, Hsu CN: **FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W635–W641.

[68] Sakharkar KR, Sakharkar MK, Chow VTK: **A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa.** *In Silico Biol* 2004, **4**(3):355–360.

[69] Smith C: **Drug target identification: a question of biology.** *Nature* 2004, **428**(6979):225–231.

[70] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.** *Nucleic Acids Res* 2005, **33**(18):5868–5877.

[71] Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S: **DNA methylation profiling of human chromosomes 6, 20 and 22.** *Nat Genet* 2006, **38**(12):1378–1385.

[72] Midorikawa Y, Yamamoto S, Ishikawa S, Kamimura N, Igarashi H, Sugimura H, Makuuchi M, Aburatani H: **Molecular karyotyping of human hepatocellular carcinoma using single-nucleotide polymorphism arrays.** *Oncogene* 2006, **25**(40):5581–5590.

[73] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**(6871):530–536.

[74] Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10**(8):789–799.

[75] Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES: **Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics.** *Proc Natl Acad Sci U S A* 2003, **100**(2):605–610.

[76] Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, Ajay, Kouros-Mehr H, Fridlyand J, Jain A, Collins C, Nishizuka S, Tonon G, Roschke A, Gehlhaus K, Kirsch I, Scudiero DA, Gray JW, Weinstein JN: **Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel.** *Mol Cancer Ther* 2006, **5**(4):853–867.

[77] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252**(5013):1651–1656.

[78] Mégy K, Audic S, Claverie JM: **Heart-specific genes revealed by expressed sequence tag (EST) sampling.** *Genome Biol* 2002, **3**(12):RESEARCH0074.

[79] Strausberg RL, Dahl CA, Klausner RD: **New opportunities for uncovering the molecular basis of cancer.** *Nat Genet* 1997, **15 Spec No**:415–416.

[80] Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7**(10):986–995.

[81] Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100**(16):9440–9445.

[82] Velculescu V, Zhang L, Vogelstein B, Kinzler K: **Serial analysis of gene expression**. *Science* 1995, **270**:484–487.

[83] Yamamoto M, Wakatsuki T, Hada A, Ryo A: **Use of serial analysis of gene expression (SAGE) technology.** *J Immunol Methods* 2001, **250**(1-2):45–66.

[84] Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: **Gene expression profiles in normal and cancer cells.** *Science* 1997, **276**(5316):1268–1272.

[85] Dias Neto E, Garcia Correa R, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva J Wilson, Zago MA, Bordin S, Costa FF, Goldman GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, de Oliveira PS, Bucher P, Jongeneel C, O'Hare MJ, Soares F, Brentani RR, Reis LF, de Souza SJ, Simpson AJ: **Shotgun sequencing of the human transcriptome with ORF expressed sequence tags**. *PNAS* 2000, **97**(7):3491–3496.

[86] Sakabe NJ, de Souza JES, Galante PAF, de Oliveira PSL, Passetti F, Brentani H, Osório EC, Zaiats AC, Leerkes MR, Kitajima JP, Brentani RR, Strausberg RL, Simpson AJG, de Souza SJ: **ORESTES are enriched in rare exon usage variants affecting the encoded proteins.** *C R Biol* 2003, **326**(10-11):979–985.

[87] Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool**. *J.Mol Biol* 1990, **215**:403–410.

[88] Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**(9):868–877.

[89] Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656–664.

[90] Strausberg RL, Camargo AA, Riggins GJ, Schaefer CF, de Souza SJ, Grouse LH, Lal A, Buetow KH, Boon K, Greenhut SF, Simpson AJG: **An international database and integrated analysis tools for the study of cancer gene expression.** *Pharmacogenomics J* 2002, **2**(3):156–164.

[91] Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource.** *Genome Res* 2000, **10**(7):1051–1060.

[92] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2005, **33**(Database issue):D39–D45.

[93] Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 2002, **99**(7):4465–4470.

[94] Yow HK, Wong JM, Chen HS, Lee CG, Davis S, Steele GD, Chen LB: **Increased mRNA expression of a laminin-binding protein in human colon carcinoma: complete sequence of a full-length cDNA encoding the protein.** *Proc Natl Acad Sci U S A* 1988, **85**(17):6394–6398.

[95] Rechreche H, Mallo GV, Montalto G, Dagorn JC, Iovanna JL: **Cloning and expression of the mRNA of human galectin-4, an S-type lectin down-regulated in colorectal cancer.** *Eur J Biochem* 1997, **248**:225–230.

[96] Kasai H, Nadano D, Hidaka E, Higuchi K, Kawakubo M, Sato TA, Nakayama J: **Differential expression of ribosomal proteins in human normal and neoplastic colorectum.** *J Histochem Cytochem* 2003, **51**(5):567–574.

[97] Amsterdam A, Sadler KC, Lai K, Farrington S, Bronson RT, Lees JA, Hopkins N: **Many ribosomal protein genes are cancer genes in zebrafish.** *PLoS Biol* 2004, **2**(5):E139.

[98] Joslin JM, Fernald AA, Tennant TR, Davis EM, Kogan SC, Anastasi J, Crispino JD, Beau MML: **Haploinsufficiency of EGR1, a candidate gene in the del(5q), leads to the development of myeloid disorders.** *Blood* 2007.

[99] Chandrasekharappa SC, Gross LA, King SE, Collins FS: **The human NME2 gene lies within 18kb of NME1 in chromosome 17.** *Genes Chromosomes Cancer* 1993, **6**(4):245–248.

[100] Wood LJ, Mukherjee M, Dolde CE, Xu Y, Maher JF, Bunton TE, Williams JB, Resar LM: **HMG-I/Y, a new c-Myc target gene and potential oncogene.** *Mol Cell Biol* 2000, **20**(15):5490–5502.

[101] Emami S, Rodrigues S, Rodrigue CM, Floch NL, Rivat C, Attoub S, Bruyneel E, Gespach C: **Trefoil factor family (TFF) peptides and cancer progression.** *Peptides* 2004, **25**(5):885–898.

[102] Wilkinson KD: **Signal transduction: aspirin, ubiquitin and cancer.** *Nature* 2003, **424**(6950):738–739.

[103] Gruber AD, Elble RC, Ji HL, Schreur KD, Fuller CM, Pauli BU: **Genomic cloning, molecular characterization, and functional analysis of human CLCA1, the first human member of the family of Ca2+-activated Cl- channel proteins.** *Genomics* 1998, **54**(2):200–214.

[104] Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA, Neto ED, Grivet M, Gruber A, Guimaraes PEM, Hide W, Iseli C, Jongeneel CV, Kelso J, Nagai MA, Ojopi EPB, Osorio EC, Reis EMR, Riggins GJ, Simpson AJG, de Souza S, Stevenson BJ, Strausberg RL, Tajara EH, Verjovski-Almeida S, Acencio ML, Bengtson MH, Bettoni F, Bodmer WF, Briones MRS, Camargo LP, Cavenee W, Cerutti JM, Andrade LEC, dos Santos PCC, Costa MCR, da Silva IT, Estécio MRH, Ferreira KS, Furnari FB, Faria M, Galante PAF, Guimaraes GS, Holanda AJ, Kimura ET, Leerkes MR, Lu X, Maciel RMB, Martins EAL, Massirer KB, Melo ASA, Mestriner CA, Miracca EC, Miranda LL, Nobrega FG, Oliveira PS, Paquola ACM, Pandolfi JRC, de Moura Campos Pardini MI, Passetti F, Quackenbush J, Schnabel B, Sogayar MC, Souza JE, Valentini SR, Zaiats AC, Amaral EJ, Arnaldi LAT, de Araújo AG, de Bessa SA, Bicknell DC, de Camaro MER, Carraro DM, Carrer H, Carvalho AF, Colin C, Costa F, Curcio C, da Silva IDCG, da Silva NP, Dellamano M, El-Dorry H, Espreafico EM, Ferreira AJS, Ferreira CA, Fortes MAHZ, Gama AH, Giannella-Neto D, Giannella MLCC, Giorgi RR, Goldman GH, Goldman MHS, Hackel C, Ho PL, Kimura EM, Kowalski LP, Krieger JE, Leite LCC, Lopes A, Luna AMSC, Mackay A, Mari SKN, Marques AA, Martins WK, Montagnini A, Neto MM, Nascimento ALTO, Neville AM, Nobrega MP, O'Hare MJ, Otsuka AY, de Melo AIR, Paco-Larson ML, Pereira GG, da Silva NP, Pesquero JB, Pessoa JG, Rahal P, Rainho CA, Rodrigues V, Rogatto SR, Romano CM, Romeiro JG, Rossi BM, Rusticci M, de Sá RG, Anna SCS, Sarmazo ML, de Lima E Silva TC, Soares FA, de Fátima Sonati M, de Freitas Sousa J, Queiroz D, Valente V, Vettore AL, Villanova FE, Zago MA, Zalcberg H, Consortium HCGPGAPA, Consortium HCGPS: **The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags.** *Proc Natl Acad Sci U S A* 2003, **100**(23):13418–13423.

[105] Staal SP: **Molecular cloning of the akt oncogene and its human homologues AKT1 and AKT2: amplification of AKT1 in a primary human gastric adenocarcinoma.** *Proc Natl Acad Sci U S A* 1987, **84**(14):5034–5037.

[106] Enomoto H, Ozaki T, Takahashi E, Nomura N, Tabata S, Takahashi H, Ohnuma N, Tanabe M, Iwai J, Yoshida H: **Identification of human DAN gene, mapping to the putative neuroblastoma tumor suppressor locus.** *Oncogene* 1994, **9**(10):2785–2791.

[107] Bauskin AR, Brown DA, Kuffner T, Johnen H, Luo XW, Hunter M, Breit SN: **Role of macrophage inhibitory cytokine-1 in tumorigenesis and diagnosis of cancer.** *Cancer Res* 2006, **66**(10):4983–4986.

[108] Workman P: **Overview: translating Hsp90 biology into Hsp90 drugs.** *Curr Cancer Drug Targets* 2003, **3**(5):297–300.

[109] McCarthy MI, Smedley D, Hide W: **New methods for finding disease-susceptibility genes: impact and potential.** *Genome Biol* 2003, **4**(10):119.

[110] de Belle I, Huang RP, Fan Y, Liu C, Mercola D, Adamson ED: **p53 and Egr-1 additively suppress transformed growth in HT1080 cells but Egr-1 counteracts p53-dependent apoptosis.** *Oncogene* 1999, **18**(24):3633–3642.

[111] Zhou Y, Luoh SM, Zhang Y, Watanabe C, Wu TD, Ostland M, Wood WI, Zhang Z: **Genome-wide Identification of Chromosomal Regions of Increased Tumor Expression by Transcriptome Analysis**. *Cancer Res* 2003, **63**(18):5781–5784.

[112] Reyal F, Stransky N, Bernard-Pierrot I, Vincent-Salomon A, de Rycke Y, Elvin P, Cassidy A, Graham A, Spraggon C, Desille Y, Fourquet A, Nos C, Pouillart P, Magdelenat H, Stoppa-Lyonnet D, Couturier J, Sigal-Zafrani B, Asselain B, Sastre-Garau X, Delattre O, Thiery JP, Radvanyi F: **Visualizing Chromosomes as Transcriptome Correlation Maps: Evidence of Chromosomal Domains Containing Co-expressed Genes–A Study of 130 Invasive Ductal Breast Carcinomas**. *Cancer Res* 2005, **65**(4):1376–1383.

[113] Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA: **Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers.** *Cell* 2004, **118**(5):555–566.

[114] Peterson CL, Laniel MA: **Histones and histone modifications.** *Curr Biol* 2004, **14**(14):R546–R551.

[115] Horn PJ, Peterson CL: **Molecular biology. Chromatin higher order folding–wrapping up transcription.** *Science* 2002, **297**(5588):1824–1827.

[116] Woodcock CL, Dimitrov S: **Higher-order structure of chromatin and chromosomes.** *Curr Opin Genet Dev* 2001, **11**(2):130–135.

[117] Cremer T, Küpper K, Dietzel S, Fakan S: **Higher order chromatin architecture in the cell nucleus: on the way from structure to function.** *Biol Cell* 2004, **96**(8):555–567.

[118] Lodén M, van Steensel B: **Whole-genome views of chromatin structure.** *Chromosome Res* 2005, **13**(3):289–298.

[119] Cremer T, Cremer C: **Chromosome territories, nuclear architecture and gene regulation in mammalian cells.** *Nat Rev Genet* 2001, **2**(4):292–301.

[120] Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA: **Differences in the localization and morphology of chromosomes in the human nucleus.** *J Cell Biol* 1999, **145**(6):1119–1131.

[121] Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S: **Chromosome territories–a functional nuclear landscape.** *Curr Opin Cell Biol* 2006, **18**(3):307–316.

[122] Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JPZ, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442**(7104):772–778.

[123] Meersseman G, Pennings S, Bradbury EM: **Mobile nucleosomes–a general behavior.** *EMBO J* 1992, **11**(8):2951–2959.

[124] Pennings S, Meersseman G, Bradbury EM: **Mobility of positioned nucleosomes on 5 S rDNA.** *J Mol Biol* 1991, **220**:101–110.

[125] Pennings S, Meersseman G, Bradbury EM: **Linker histones H1 and H5 prevent the mobility of positioned nucleosomes.** *Proc Natl Acad Sci U S A* 1994, **91**(22):10275–10279.

[126] Imbalzano AN, Kwon H, Green MR, Kingston RE: **Facilitated binding of TATA-binding protein to nucleosomal DNA.** *Nature* 1994, **370**(6489):481–485.

[127] Santos-Rosa H, Caldas C: **Chromatin modifier enzymes, the histone code and cancer.** *Eur J Cancer* 2005, **41**(16):2381–2402.

[128] Marmorstein R: **Protein modules that manipulate histone tails for chromatin regulation.** *Nat Rev Mol Cell Biol* 2001, **2**(6):422–432.

[129] Lund AH, van Lohuizen M: **Epigenetics and cancer**. *Genes Dev.* 2004, **18**(19):2315–2335.

[130] Villar-Garea A, Esteller M: **Histone deacetylase inhibitors: understanding a new wave of anticancer agents**. *Int.J.Cancer 2004.Nov.1;112.(2):171.-8.* 2004, **112**:171–178.

[131] Francis NJ, Kingston RE, Woodcock CL: **Chromatin Compaction by a Polycomb Group Protein Complex**. *Science* 2004, **306**(5701):1574–1577.

[132] Valk-Lingbeek ME, Bruggeman SW, van Lohuizen M: **Stem Cells and Cancer: The Polycomb Connection**. *Cell* 2004, **118**(4):409–418.

[133] Cremer M, Küpper K, Wagler B, Wizelman L, von Hase J, Weiland Y, Kreja L, Diebold J, Speicher MR, Cremer T: **Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei.** *J Cell Biol* 2003, **162**(5):809–820.

[134] Lenburg M, Liou L, Gerry N, Frampton G, Cohen H, Christman M: **Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data**. *BMC Cancer* 2003, **3**:31–.

[135] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *PNAS* 2001, **98**(24):13790–13795.

[136] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP: **Gene expression correlates of clinical prostate cancer behavior**. *Cancer Cell* 2002, **1**(2):203–209.

[137] Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection**. *PNAS* 2001, **98**:31–36.

[138] Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, Tsai CJ, Zhang S: **Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes**. *BMC Bioinformatics* 2004, **5**:81–.

[139] Su A, Wiltshire T, Batalov S, Lapp H, Ching K, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke M, Walker J, Hogenesch J: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc.Natl.Acad.Sci.U.S.A* 2004, **101**(16):6062–6067.

[140] Ficarra V, Martignoni G, Maffei N, Brunelli M, Novara G, Zanolla L, Pea M, Artibani W: **Original and reviewed nuclear grading according to the Fuhrman system: a multivariate analysis of 388 patients with conventional renal cell carcinoma**. *Cancer 2005.Jan.1;103.(1):68.-75.* 2005, **103**:68–75.

[141] Lohse C, Blute M, Zincke H, Weaver A, Cheville J: **Comparison of standardized and nonstandardized nuclear grade of renal cell carcinoma to predict outcome among 2,042 patients**. *Am.J.Clin.Pathol.2002.Dec.;118.(6.):877.-86.* 2002, **118**:877–886.

[142] Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L, Neri A, Califano A, Dalla-Favera R: **Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells.** *J Exp Med* 2001, **194**(11):1625–1638.

[143] Geller SC, Gregg JP, Hagerman P, Rocke DM: **Transformation and normalization of oligonucleotide microarray data.** *Bioinformatics* 2003, **19**(14):1817–1823.

[144] Michel L, Benezra R, Diaz-Rodriguez E: **MAD2 dependent mitotic checkpoint defects in tumorigenesis and tumor cell death: a double edged sword**. *Cell Cycle 2004.Aug.;3(8.):990.-2.Epub.2004.Aug.25.* 2004, **3**:990–992.

[145] Ruggero D, Montanaro L, Ma L, Xu W, Londei P, Cordon-Cardo C, Pandolfi PP: **The translation factor eIF-4E promotes tumor formation and cooperates with c-Myc in lymphomagenesis**. *Nat Med* 2004, **10**(5):484–486.

[146] Huminiecki L, Lloyd A, Wolfe K: **Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases**. *BMC Genomics* 2003, **4**:31–.

[147] Lercher M, Urrutia A, Hurst L: **Clustering of housekeeping genes provides a unified model of gene order in the human genome**. *Nat Genet.2002.Jun.;31.(2):180.-3.Epub.2002.May.6.* 2002, **31**:180–183.

[148] Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD: **A unification of mosaic structures in the human genome**. *Hum.Mol.Genet.* 2003, **12**(19):2411–2415.

[149] Gazave E, Gautier P, Gilchrist S, Bickmore WA: **Does radial nuclear organisation influence DNA damage?** *Chromosome Res* 2005, **13**(4):377–388.

[150] Arndt PF, Hwa T, Petrov DA: **Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects.** *J Mol Evol* 2005, **60**(6):748–763.

[151] Chuang JH, Li H: **Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome.** *PLoS Biol* 2004, **2**(2):E29.

[152] Williams EJ, Hurst LD: **The proteins of linked genes evolve at similar rates**. *Nature* 2000, **407**(6806):900–903.

[153] Lercher M, Chamary J, Hurst L: **Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile**. *Genome Res.2004.Jun.;14(6.):1002.-13.* 2004, **14**:1002–1013.

[154] Sémon M, Duret L: **Evolutionary origin and maintenance of coexpressed gene clusters in mammals.** *Mol Biol Evol* 2006, **23**(9):1715–1723.

[155] Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucl.Acids Res.* 2004, **32**(5):1792–1797.

[156] Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.

[157] Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite**. *Trends in Genetics* 2000, **16**(6):276–277.

[158] Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13**(5):555–556.

[159] Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG: **Natural selection on protein-coding genes in the human genome.** *Nature* 2005, **437**(7062):1153–1157.

[160] Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci U S A* 2003, **100**(20):11484–11489.

[161] Sequencing C, Consortium A: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437**(7055):69–87.

[162] Karolchik D, Baertsch R, Diekhans M, Furey T, Hinrichs A, Lu Y, Roskin K, Schwartz M, Sugnet C, Thomas D, Weber R, Haussler D, Kent W: **The UCSC Genome Browser Database**. *Nucl.Acids Res.* 2003, **31**:51–54.

[163] Takai D, Jones PA: **Comprehensive analysis of CpG islands in human chromosomes 21 and 22**. *PNAS* 2002, **99**(6):3740–3745.

[164] Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7**(2):97–104.

[165] Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, Jong PJD, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Albà MM, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hübner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, López-Otín C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F, Consortium

RGSP: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428**(6982):493–521.

[166] Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM: **Heterotachy in mammalian promoter evolution.** *PLoS Genet* 2006, **2**(4):e30.

[167] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Etten WJV, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, Group ISMW: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409**(6822):928–933.

[168] Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive identification of exonic splicing enhancers in human genes.** *Science* 2002, **297**(5583):1007–1013.

[169] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**(5576):2225–2229.

[170] Consortium TIH: **The International HapMap Project.** *Nature* 2003, **426**(6968):789–796.

[171] Sadoni N, Langer S, Fauth C, Bernardi G, Cremer T, Turner BM, Zink D: **Nuclear organization of mammalian genomes. Polar chromosome territories build up functionally distinct higher order compartments.** *J Cell Biol* 1999, **146**(6):1211–1226.

[172] Hardison RC, Roskin KM, Yang S, Diekhans M, Kent W, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey TS, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F, Haussler D: **Covariation in Frequencies**

of Substitution, Deletion, Transposition, and Recombination During Eutherian Evolution. *Genome Res.* 2003, **13**:13–26.

[173] Lercher MJ, Hurst LD: **Human SNP variability and mutation rate are higher in regions of high recombination.** *Trends Genet* 2002, **18**(7):337–340.

[174] Hosack D, Dennis G, Sherman B, Lane H, Lempicki R: **Identifying biological themes within lists of genes with EASE**. *Genome Biology* 2003, **4**(10):R70–.

[175] Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev Genet* 2006, **7**(2):98–108.

[176] Castresana J: **Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content.** *Nucleic Acids Res* 2002, **30**(8):1751–1756.

[177] Zhang L, Li W: **Mammalian housekeeping genes evolve more slowly than tissue-specific genes**. *Mol Biol Evol* 2004, **21**(2):236–239.

[178] Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci U S A* 1993, **90**(24):11995–11999.

[179] Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, Gingeras TR, Schreiber SL, Lander ES: **Genomic maps and comparative analysis of histone modifications in human and mouse.** *Cell* 2005, **120**(2):169–181.

[180] McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature* 1991, **351**(6328):652–654.

[181] dos Reis M, Savva R, Wernisch L: **Solving the riddle of codon usage preferences: a test for translational selection.** *Nucleic Acids Res* 2004, **32**(17):5036–5044.

[182] Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.** *Genome Biol* 2005, **6**(9):R75.

[183] Consortium TIH: **A haplotype map of the human genome.** *Nature* 2005, **437**(7063):1299–1320.

[184] McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome.** *Science* 2004, **304**(5670):581–584.

[185] Surrallés J, Ramírez MJ, Marcos R, Natarajan AT, Mullenders LHF: **Clusters of transcription-coupled repair in the human genome.** *Proc Natl Acad Sci U S A* 2002, **99**(16):10571–10574.

[186] Gérard A, Polo SE, Roche D, Almouzni G: **Methods for studying chromatin assembly coupled to DNA repair.** *Methods Enzymol* 2006, **409**:358–374.

[187] Loizou JI, Murr R, Finkbeiner MG, Sawan C, Wang ZQ, Herceg Z: **Epigenetic information in chromatin: the code of entry for DNA repair.** *Cell Cycle* 2006, **5**(7):696–701.

[188] Rubbi CP, Milner J: **p53 is a chromatin accessibility factor for nucleotide excision repair of DNA damage.** *EMBO J* 2003, **22**(4):975–986.

[189] Vinogradov AE: **Noncoding DNA, isochores and gene expression: nucleosome formation potential.** *Nucleic Acids Res* 2005, **33**(2):559–563.

[190] of colorectal cancer G: **http://www.cancer.gov/cancertopics/pdq/ genetics/colorectal/healthprofessional**.

[191] Turnbull CL: **Sequence stability of the APC gene: the role of DNA repair mechanisms in colon carcinogenesis**. *PhD thesis*, University of Edinburgh 2007.

[192] Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease**. *Nat Genet 2003.Mar.;33.Suppl:228.-37.* 2003, **33 Suppl**:228–237.

[193] Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**(5281):1516–1517.

[194] Ye Z, Song H, Higgins JPT, Pharoah P, Danesh J: **Five glutathione s-transferase gene variants in 23,452 cases of lung cancer and 30,397 controls: meta-analysis of 130 studies.** *PLoS Med* 2006, **3**(4):e91.

[195] Ackerman H, Usen S, Jallow M, Sisay-Joof F, Pinder M, Kwiatkowski DP: **A comparison of case-control and family-based association methods: the example of sickle-cell and malaria.** *Ann Hum Genet* 2005, **69**(Pt 5):559–565.

[196] Begovich AB, Carlton VEH, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM, Conn MT, Chang M, Chang SYP, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F, Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ, Gregersen PK: **A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis.** *Am J Hum Genet* 2004, **75**(2):330–337.

[197] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration.** *Science* 2005, **308**(5720):385–389.

[198] Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA: **Complement factor H polymorphism and age-related macular degeneration.** *Science* 2005, **308**(5720):421–424.

[199] Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA: **Complement factor H variant increases the risk of age-related macular degeneration.** *Science* 2005, **308**(5720):419–421.

[200] Haenszel W, Kurihara M: **Studies of Japanese migrants. I. Mortality from cancer and other diseases among Japanese in the United States.** *J Natl Cancer Inst* 1968, **40**:43–68.

[201] Wood RD, Mitchell M, Lindahl T: **Human DNA repair genes, 2005.** *Mutat Res* 2005, **577**(1-2):275–283.

[202] White JA, McAlpine PJ, Antonarakis S, Cann H, Eppig JT, Frazer K, Frezal J, Lancet D, Nahmias J, Pearson P, Peters J, Scott A, Scott H, Spurr N, Talbot C, Povey S: **Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee.** *Genomics* 1997, **45**(2):468–471.

[203] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29.

[204] de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies.** *Nat Genet* 2005, **37**(11):1217–1223.

[205] Tenesa A, Dunlop MG: **Validity of tagging SNPs across populations for association studies.** *Eur J Hum Genet* 2006, **14**(3):357–363.

[206] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**(3):559–575.

[207] Yue P, Melamud E, Moult J: **SNPs3D: candidate gene and SNP selection for association studies.** *BMC Bioinformatics* 2006, **7**:166.

[208] Zhu Y, Spitz MR, Amos CI, Lin J, Schabath MB, Wu X: **An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology.** *Cancer Res* 2004, **64**(6):2251–2257.

[209] Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes.** *Nat Genet* 2007, **39**(7):906–913.

[210] Hurst LD, Pál C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5**(4):299–310.

[211] Pál C, Hurst LD: **Evidence for co-evolution of gene order and recombination rate.** *Nat Genet* 2003, **33**(3):392–395.

[212] Nei M: **Modification of linkage intensity by natural selection.** *Genetics* 1967, **57**(3):625–641.

[213] Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K: **Evidence of a large-scale functional organization of mammalian chromosomes.** *PLoS Genet* 2005, **1**(3):e33.

[214] Sekiguchi J, Ferguson DO, Chen HT, Yang EM, Earle J, Frank K, Whitlow S, Gu Y, Xu Y, Nussenzweig A, Alt FW: **Genetic interactions between ATM and the nonhomologous end-joining factors in genomic stability and development.** *Proc Natl Acad Sci U S A* 2001, **98**(6):3243–3248.

[215] Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucl.Acids Res.* 2002, **30**(17):3894–3900.

[216] Wang Z, Moult J: **SNPs, protein structure, and disease**. *Hum.Mutat.2001.Apr;17(4):263.-70.* 2001, **17**:263–270.

[217] Clifford RJ, Edmonson MN, Nguyen C, Buetow KH: **Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms.** *Bioinformatics* 2004, **20**(7):1006–1014.

[218] Ng PC, Henikoff S: **Accounting for Human Polymorphisms Predicted to Affect Protein Function**. *Genome Res.* 2002, **12**(3):436–446.

[219] Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian I, Haussler D: **Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology**. *Comput.Appl.Biosci.* 1996, **12**(4):327–345.

[220] Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN: **PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.** *Protein Eng* 1999, **12**(5):387–394.

[221] Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78–94.

[222] Burge CB, Karlin S: **Finding the genes in genomic DNA.** *Curr Opin Struct Biol* 1998, **8**(3):346–354.

[223] Farrington SM, Tenesa A, Barnetson R, Wiltshire A, Prendergast J, Porteous M, Campbell H, Dunlop MG: **Germline susceptibility to colorectal cancer due to base-excision repair gene defects.** *Am J Hum Genet* 2005, **77**:112–119.

[224] Futreal P, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nat Rev Cancer* 2004, **4**(3):177–183.

[225] Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, Albers-Akkers MT, Marcos JGI, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K: **Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24.** *Nat Genet* 2007, **39**(5):631–637.

[226] Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF, Hoover R, Hunter DJ, Chanock SJ, Thomas G: **Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.** *Nat Genet* 2007, **39**(5):645–649.

[227] Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, Greenway SC, Stram DO, Marchand LL, Kolonel LN, Frasco M, Wong D, Pooler LC, Ardlie K, Oakley-Girvan I, Whittemore AS, Cooney KA, John EM, Ingles SA, Altshuler

D, Henderson BE, Reich D: **Multiple regions within 8q24 independently affect risk for prostate cancer.** *Nat Genet* 2007, **39**(5):638–644.

[228] Witte JS: **Multiple prostate cancer risk variants on 8q24.** *Nat Genet* 2007, **39**(5):579–580.

[229] Schumacher FR, Feigelson HS, Cox DG, Haiman CA, Albanes D, Buring J, Calle EE, Chanock SJ, Colditz GA, Diver WR, Dunning AM, Freedman ML, Gaziano JM, Giovannucci E, Hankinson SE, Hayes RB, Henderson BE, Hoover RN, Kaaks R, Key T, Kolonel LN, Kraft P, Marchand LL, Ma J, Pike MC, Riboli E, Stampfer MJ, Stram DO, Thomas G, Thun MJ, Travis R, Virtamo J, Andriole G, Gelmann E, Willett WC, Hunter DJ: **A common 8q24 variant in prostate and breast cancer from a large nested case-control study.** *Cancer Res* 2007, **67**(7):2951–2956.

[230] Consortium WTCC: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**(7145):661–678.

[231] Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous MEM, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellié C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG: **Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24.** *Nat Genet* 2007.

[232] Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K, Consortium CORGI, Silver A, Atkin W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R: **A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21.** *Nat Genet* 2007, **39**(8):984–988.

[233] Byrne JA, Tomasetto C, Garnier JM, Rouyer N, Mattei MG, Bellocq JP, Rio MC, Basset P: **A screening method to identify genes commonly overexpressed in carcinomas and the identification of a novel complementary DNA sequence.** *Cancer Res* 1995, **55**(13):2896–2903.

[234] Byrne JA, Nourse CR, Basset P, Gunning P: **Identification of homo- and heteromeric interactions between members of the breast carcinoma-associated D52 protein family using the yeast two-hybrid system.** *Oncogene* 1998, **16**(7):873–881.

[235] Mori T, Li Y, Hata H, Ono K, Kochi H: **NIRF, a novel RING finger protein, is involved in cell-cycle regulation.** *Biochem Biophys Res Commun* 2002, **296**(3):530–536.

[236] Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS: **Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).** *Genome Res* 2006, **16**:123–131.

[237] Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128**(6):1231–1245.

[238] Knuutila S, Aalto Y, Autio K, Björkqvist AM, El-Rifai W, Hemmer S, Huhta T, Kettunen E, Kiuru-Kuhlefelt S, Larramendy ML, Lushnikova T, Monni O, Pere H, Tapper J, Tarkkanen M, Varis A, Wasenius VM, Wolf M, Zhu Y: **DNA copy number losses in human neoplasms.** *Am J Pathol* 1999, **155**(3):683–694.

[239] Joos S, Küpper M, Ohl S, von Bonin F, Mechtersheimer G, Bentz M, Marynen P, Möller P, Pfreundschuh M, Trümper L, Lichter P: **Genomic imbalances including amplification of the tyrosine kinase gene JAK2 in CD30+ Hodgkin cells.** *Cancer Res* 2000, **60**(3):549–552.

[240] Daniely M, Aviram A, Adams EF, Buchfelder M, Barkai G, Fahlbusch R, Goldman B, Friedman E: **Comparative genomic hybridization analy-**

sis of nonfunctioning pituitary tumors. *J Clin Endocrinol Metab* 1998, **83**(5):1801–1805.

[241] Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, Jakobsdottir M, Kostic J, Magnusdottir DN, Ghosh S, Agnarsson K, Birgisdottir B, Roux LL, Olafsdottir A, Blondal T, Andresdottir M, Gretarsdottir OS, Bergthorsson JT, Gudbjartsson D, Gylfason A, Thorleifsson G, Manolescu A, Kristjansson K, Geirsson G, Isaksson H, Douglas J, Johansson JE, Bälter K, Wiklund F, Montie JE, Yu X, Suarez BK, Ober C, Cooney KA, Gronberg H, Catalona WJ, Einarsson GV, Barkardottir RB, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K: **A common variant associated with prostate cancer in European and African populations.** *Nat Genet* 2006, **38**(6):652–658.

[242] Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: **Emergence of young human genes after a burst of retroposition in primates.** *PLoS Biol* 2005, **3**(11):e357.

[243] Takeda J, Seino S, Bell GI: **Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues.** *Nucleic Acids Res* 1992, **20**(17):4613–4620.

[244] Dryja TP, Mukai S, Petersen R, Rapaport JM, Walton D, Yandell DW: **Parental origin of mutations of the retinoblastoma gene.** *Nature* 1989, **339**(6225):556–558.

[245] Stransky N, Vallot C, Reyal F, Bernard-Pierrot I, de Medina SGD, Segraves R, de Rycke Y, Elvin P, Cassidy A, Spraggon C, Graham A, Southgate J, Asselain B, Allory Y, Abbou CC, Albertson DG, Thiery JP, Chopin DK, Pinkel D, Radvanyi F: **Regional copy number-independent deregulation of transcription in cancer.** *Nat Genet* 2006, **38**(12):1386–1396.

[246] Dumont P, Leu JIJ, Pietra ACD, George DL, Murphy M: **The codon 72 polymorphic variants of p53 have markedly different apoptotic potential.** *Nat Genet* 2003, **33**(3):357–365.

[247] Jamroziak K, M?ynarski W, Balcerczak E, Mistygacz M, Trelinska J, Mirowski M, Bodalski J, Robak T: **Functional C3435T polymorphism of MDR1 gene: an impact on genetic susceptibility and clinical outcome of childhood acute lymphoblastic leukemia.** *Eur J Haematol* 2004, **72**(5):314–321.

[248] Wu Y, Berends MJ, Post JG, Mensink RG, Verlind E, Sluis TVD, Kempinga C, Sijmons RH, van der Zee AG, Hollema H, Kleibeuker JH, Buys CH, Hofstra RM: **Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms.** *Gastroenterology* 2001, **120**(7):1580–1587.

[249] Kiyohara C: **Genetic polymorphism of enzymes involved in xenobiotic metabolism and the risk of colorectal cancer.** *J Epidemiol* 2000, **10**(5):349–360.

[250] Harrison DJ, Hubbard AL, MacMillan J, Wyllie AH, Smith CA: **Microsomal epoxide hydrolase gene polymorphism and susceptibility to colon cancer.** *Br J Cancer* 1999, **79**:168–171.

[251] El-Omar EM, Carrington M, Chow WH, McColl KE, Bream JH, Young HA, Herrera J, Lissowska J, Yuan CC, Rothman N, Lanyon G, Martin M, Fraumeni JF, Rabkin CS: **Interleukin-1 polymorphisms associated with increased risk of gastric cancer.** *Nature* 2000, **404**(6776):398–402.

[252] Strassburg CP, Vogel A, Kneip S, Tukey RH, Manns MP: **Polymorphisms of the human UDP-glucuronosyltransferase (UGT) 1A7 gene in colorectal cancer.** *Gut* 2002, **50**(6):851–856.

[253] Yeh CC, Sung FC, Tang R, Chang-Chieh CR, Hsieh LL: **Polymorphisms of the XRCC1, XRCC3, & XPD genes, and colorectal cancer risk: a case-control study in Taiwan.** *BMC Cancer* 2005, **5**:12.

[254] Porter TR, Richards FM, Houlston RS, Evans DGR, Jankowski JA, Macdonald F, Norbury G, Payne SJ, Fisher SA, Tomlinson I, Maher ER: **Contribution of cyclin d1 (CCND1) and E-cadherin (CDH1) polymorphisms to familial and sporadic colorectal cancer.** *Oncogene* 2002, **21**(12):1928–1933.

[255] Orimo H, Nakajima E, Yamamoto M, Ikejima M, Emi M, Shimada T: **Association between single nucleotide polymorphisms in the hMSH3 gene and sporadic colon cancer with microsatellite instability.** *J Hum Genet* 2000, **45**(4):228–230.

[256] de Jong MM, Hofstra RMW, Kooi KA, Westra JL, Berends MJW, Wu Y, Hollema H, van der Sluis T, van der Graaf WTA, de Vries EGE, Schaapveld M, Sijmons RH, te Meerman GJ, Kleibeuker JH: **No association between two MLH3 variants (S845G and P844L)and colorectal cancer risk.** *Cancer Genet Cytogenet* 2004, **152**:70–71.

[257] Yoshimura K, Hanaoka T, Ohnami S, Ohnami S, Kohno T, Liu Y, Yoshida T, Sakamoto H, Tsugane S: **Allele frequencies of single nucleotide polymorphisms (SNPs) in 40 candidate genes for gene-environment studies on cancer: data from population-based Japanese random samples.** *J Hum Genet* 2003, **48**(12):654–658.

[258] Shen H, Xu Y, Qian Y, Yu R, Qin Y, Zhou L, Wang X, Spitz MR, Wei Q: **Polymorphisms of the DNA repair gene XRCC1 and risk of gastric cancer in a Chinese population.** *Int J Cancer* 2000, **88**(4):601–606.

[259] Brockton N, Little J, Sharp L, Cotton SC: **N-acetyltransferase polymorphisms and colorectal cancer: a HuGE review.** *Am J Epidemiol* 2000, **151**(9):846–861.

[260] Bagnoli S, Putignano AL, Melean G, Baglioni S, Sestini R, Milla M, d'Albasio G, Genuardi M, Pacini F, Trallori G, Papi L: **Susceptibility to refractory ulcerative colitis is associated with polymorphism in the hMLH1 mismatch repair gene.** *Inflamm Bowel Dis* 2004, **10**(6):705–708.

[261] Slattery ML, Samowitz W, Curtin K, Ma KN, Hoffman M, Caan B, Neuhausen S: **Associations among IRS1, IRS2, IGF1, and IGFBP3 genetic polymorphisms and colorectal cancer.** *Cancer Epidemiol Biomarkers Prev* 2004, **13**(7):1206–1214.

[262] Landi S, Moreno V, Gioia-Patricola L, Guino E, Navarro M, de Oca J, Capella G, Canzian F, Group BCCS: **Association of common polymorphisms in**

inflammatory genes interleukin (IL)6, IL8, tumor necrosis factor alpha, NFKB1, and peroxisome proliferator-activated receptor gamma with colorectal cancer. *Cancer Res* 2003, **63**(13):3560–3566.

[263] Cahill DP, Lengauer C, Yu J, Riggins GJ, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B: **Mutations of mitotic checkpoint genes in human cancers.** *Nature* 1998, **392**(6673):300–303.

[264] Hampe J, Grebe J, Nikolaus S, Solberg C, Croucher PJP, Mascheretti S, Jahnsen J, Moum B, Klump B, Krawczak M, Mirza MM, Foelsch UR, Vatn M, Schreiber S: **Association of NOD2 (CARD 15) genotype with clinical course of Crohn's disease: a cohort study.** *Lancet* 2002, **359**(9318):1661–1665.

[265] Ewart-Toland A, Briassouli P, de Koning JP, Mao JH, Yuan J, Chan F, MacCarthy-Morrogh L, Ponder BAJ, Nagase H, Burn J, Ball S, Almeida M, Linardopoulos S, Balmain A: **Identification of Stk6/STK15 as a candidate low-penetrance tumor-susceptibility gene in mouse and human.** *Nat Genet* 2003, **34**(4):403–412.

[266] Kim HC, Wheeler JM, Kim JC, Ilyas M, Beck NE, Kim BS, Park KC, Bodmer WF: **The E-cadherin gene (CDH1) variants T340A and L599V in gastric and colorectal cancer patients in Korea.** *Gut* 2000, **47**(2):262–267.

[267] Kolodner RD, Hall NR, Lipford J, Kane MF, Rao MR, Morrison P, Wirth L, Finan PJ, Burn J, Chapman P: **Human mismatch repair genes and their association with hereditary non-polyposis colon cancer.** *Cold Spring Harb Symp Quant Biol* 1994, **59**:331–338.

[268] Krupa R, Blasiak J: **An association of polymorphism of DNA repair genes XRCC1 and XRCC3 with colorectal cancer.** *J Exp Clin Cancer Res* 2004, **23**(2):285–294.

[269] Cox DG, Pontes C, Guino E, Navarro M, Osorio A, Canzian F, Moreno V, Group BCCS: **Polymorphisms in prostaglandin synthase 2/cyclooxygenase 2 (PTGS2/COX2) and risk of colorectal cancer.** *Br J Cancer* 2004, **91**(2):339–343.

[270] Kong S, Amos CI, Luthra R, Lynch PM, Levin B, Frazier ML: **Effects of cyclin D1 polymorphism on age of onset of hereditary nonpolyposis colorectal cancer.** *Cancer Res* 2000, **60**(2):249–252.

nature genetics

Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24

Brent W Zanke[1–3], Celia MT Greenwood[1,4,5], Jagadish Rangrej[4], Rafal Kustra[1,5], Albert Tenesa[6], Susan M Farrington[6], James Prendergast[6], Sylviane Olschwang[7], Theodore Chiang[4], Edgar Crowdy[4], Vincent Ferretti[8], Philippe Laflamme[8], Saravanan Sundararajan[8], Stéphanie Roumy[8], Jean-François Olivier[8], Frédérick Robidoux[8], Robert Sladek[8], Alexandre Montpetit[8], Peter Campbell[9], Stephane Bezieau[10], Anne Marie O'Shea[9], George Zogopoulos[9], Michelle Cotterchio[1,5], Polly Newcomb[11], John McLaughlin[1,9], Ban Younghusband[12], Roger Green[12], Jane Green[12], Mary E M Porteous[13], Harry Campbell[6,14], Helene Blanche[15], Mourad Sahbatou[15], Emmanuel Tubacher[15], Catherine Bonaiti-Pellié[16], Bruno Buecher[10], Elio Riboli[17], Sebastien Kury[10], Stephen J Chanock[18], John Potter[11], Gilles Thomas[19], Steven Gallinger[1,9], Thomas J Hudson[2,8] & Malcolm G Dunlop[6]

Using a multistage genetic association approach comprising 7,480 affected individuals and 7,779 controls, we identified markers in chromosomal region 8q24 associated with colorectal cancer. In stage 1, we genotyped 99,632 SNPs in 1,257 affected individuals and 1,336 controls from Ontario. In stages 2–4, we performed serial replication studies using 4,024 affected individuals and 4,042 controls from Seattle, Newfoundland and Scotland. We identified one locus on chromosome 8q24 and another on 9p24 having combined odds ratios (OR) for stages 1–4 of 1.18 (trend; $P = 1.41 \times 10^{-8}$) and 1.14 (trend; $P = 1.32 \times 10^{-5}$), respectively. Additional analyses in 2,199 affected individuals and 2,401 controls from France and Europe supported the association at the 8q24 locus (OR = 1.16, trend; 95% confidence interval (c.i.): 1.07–1.26; $P = 5.05 \times 10^{-4}$). A summary across all seven studies at the 8q24 locus was highly significant (OR = 1.17, c.i.: 1.12–1.23; $P = 3.16 \times 10^{-11}$). This locus has also been implicated in prostate cancer[1–3].

Colorectal cancer is a common cause of cancer death in developed countries. Genetic susceptibility accounts for 35% of disease etiology[4], most of which remains to be explained. We studied 1,257 affected individuals and 1,336 matched community controls (identified by population-based random telephone dialing) provided by the Ontario Familial Colorectal Cancer Registry (OFCCR)[5]. Affected individuals with known germline *APC*, *MSH2*, *MLH1*, *MSH6* or biallelic *MUTYH* mutations were excluded. The initial screen involved three SNP panels tested using three genotyping technologies: (i) 1,536 SNPs in 227 genes implicated in DNA repair, genome instability, folate metabolism and other pathways implicated in carcinogenesis, which we genotyped successfully on 1,226 affected individuals and 1,239 controls using the Illumina GoldenGate technology[6]; (ii) the Affymetrix Gene Chip 10K coding SNP (cSNP) array containing approximately 9,701 coding nonsynonymous SNPs, which we used to genotype 1,135 affected individuals and 1,157 controls and (iii) two Affymetrix Mendel arrays each containing approximately 50,000 genomic SNPs, which we used to genotype 960 affected individuals

[1]Cancer Care Ontario, 620 University Avenue, Toronto, Ontario M5G 1L7, Canada. [2]The Ontario Institute for Cancer Research, 101 College St., Toronto M5G 2L7, Canada. [3]The University of Ottawa, Faculty of Medicine, Division of Hematology, 501 Smythe Road, Ottawa K1H 8L6, Canada. [4]Genetics and Genome Biology, Hospital for Sick Children, 15-703 TMDT East, 101 College Street, Toronto, Ontario M5G 1L7, Canada. [5]University of Toronto, Department of Public Health Sciences Health Sciences Building, 155 College Street, Toronto M5T 3M7, Canada. [6]Colon Cancer Genetics Group, University of Edinburgh Cancer Research Centre and UK Medical Research Council (MRC) Human Genetics Unit, Western General Hospital, Edinburgh UK EH4 2XU, UK. [7]INSERM U599, Institut Paoli Calmettes, F-13009 Marseille, France. [8]The McGill University and Genome Quebec Innovation Centre, 700 Dr. Penfield Ave., Montreal, Quebec H3G 1A4, Canada. [9]Samuel Lunenfeld Research Institute, Mount Sinai Hospital and University of Toronto, 600 University Ave., Toronto, Ontario M5G 1X5, Canada. [10]EA3823 and Institut des Maladies de l'Appareil Digestif (IMAD), Centre Hospitalier Universitaire Hotel-Dieu, 44093 Nantes Cedex 01, France. [11]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [12]Memorial University of Newfoundland, St. John's, Newfoundland A1C 5S7, Canada. [13]Clinical Genetics Department, University of Edinburgh, Edinburgh EH4 2XU, UK. [14]Public Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK. [15]Centre d'Etude du Polymorphisme Humain, 27 rue Juliette Dodu, F-75010 Paris, France. [16]Institut National de la Santé et de la Recherche Médicale (INSERM) U535, Hopital Paul Brousse, BP1000, 94800 Villejuif, France. [17]Faculty of Medicine, Imperial College W2 1PG, London, UK. [18]Center for Cancer Research, National Cancer Institute (NCI), US National Institutes of Health (NIH), Department of Health and Human Services (DHHS), and Division of Cancer Genetics and Epidemiology, NCI, NIH, DHHS, 8717 Grovemont Circle, Gaithersburg, Maryland 20877, USA. [19]Division of Cancer Epidemiology and Genetics National Cancer Institute 8717 Grovemont Circle, Bethesda, Maryland 20892-4605, USA. Correspondence should be addressed to T.J.H. (tom.hudson@oicr.on.ca).

Received 20 April; accepted 1 June; published online 8 July 2007; doi:10.1038/ng2089

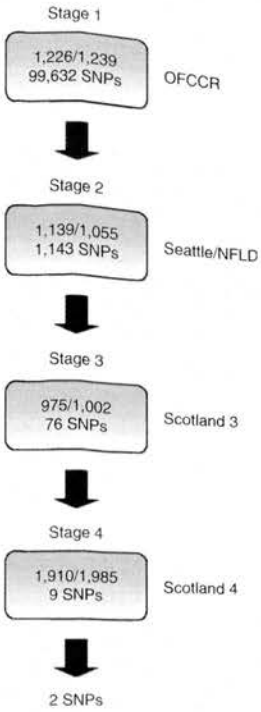


Figure 1 Study scheme. The experimental stage 1 and three sequential validation stages (2–4) are illustrated, with the number of successfully genotyped affected individuals and controls and the number of analyzed markers. For OFCCR, the case-control numbers refer to the Illumina custom array. Attrition of candidate markers of colorectal cancer is indicated, with two SNPs remaining significantly associated with disease after validation testing.

and 984 controls. After excluding SNPs with low allele frequency or unreliable genotyping data (see Methods for details), we compared 99,632 SNPs among affected individuals and controls (**Supplementary Fig. 1** online).

To determine which associations identified in Stage 1 were true, we performed sequential replication studies in independent case-control series (**Fig. 1**). The first replication panel tested 1,143 markers selected from the stage 1 SNP panels. We used different *P* value thresholds for marker selection from the three Ontario genotyping data sets owing to *a priori* considerations about the number of 'true' associations in each data set[7] (see Methods for details). Our approach for marker selection was based on optimizing the stratified false discovery rate (FDR). We estimated FDR to be 0.68, reduced from an estimate of 0.94 obtained by a naive selection based on *P* values alone. We then tested the resultant SNP panel in two case-control sets, one from Seattle (687/688) and another from Newfoundland (452/367). To maximize power yet allow for heterogeneity between these two populations, disease-marker associations were estimated using a random effect model for the log odds ratios[8]. Replication was considered to have occurred when the significance was <0.10 in both Ontario and the combined validation data sets under the same model (that is, dominant, recessive or trend) and when the association was in the same direction. We identified 76 markers that achieved this threshold (**Supplementary Table 1** online).

We then investigated the 76 putative associations identified in stages 1 and 2 in an early-onset colorectal cancer case-control set from Scotland (mean age: affected individuals, 49.1 years; controls, 50.9 years). We generated genotypes (975 affected individuals, 1,002 controls) for each of the 76 SNPs either directly ($n = 28$) or through a tagging strategy ($n = 41$) ($r^2 > 0.7$) using data from a concurrent genome-wide scan in this case-control set using Illumina HumanHap300 and HumanHap240S arrays genotyped on the Infinium platform. The seven SNPs that were neither tagged nor genotyped directly by this approach were genotyped using Applied Biosystems (ABI) TaqMan in the same Scottish case-control sample set (see **Supplementary Table 2** online for all genotyping). We then assessed all 76 putative associations from stages 1 and 2 using an allelic model of inheritance for evidence of association between each marker or tag(s) and colorectal cancer phenotype. We used a χ^2 test of allelic counts (one degree of freedom) unless two or more tagging SNPs were needed, which required a χ^2 test using estimated counts of the tagging haplotype, as implemented in Haploview[2]. Using a cut-off threshold of $P < 0.10$, we found nine SNPs showing evidence for replication in the Scottish stage 3 sample set (*P* values between 0.0025–0.0736; **Table 1**). These nine loci were then tested for association (stage 4) in a further, independent Scottish case-control series of older-onset cases (mean age: affected individuals, 65.8 years; controls 67.9 years). Excluding missing values, we generated TaqMan genotypes for 1,910 affected individuals and 1,985 controls for the corresponding SNP identified in stages 1 and 2 (but not the tags used in stage 3). Two of the associations identified in stage 3 were replicated further in this stage 4 Scottish case-control sample set: rs10505477 (tagged by rs6983267 in stage 3, $P = 0.03$) and rs719725 (tagged by rs7857628 or rs206636213 in stage 3, $P = 0.0025$). Stage 4 associations detected for rs10505477 gave an OR of 1.16 (c.i. 1.11–1.21; $P = 0.001$), and for rs719725, OR = 1.10 (c.i. 1.05–1.15; $P = 0.037$) in an allelic model. **Supplementary Table 2** contains data for these 76 markers from all four stages.

To address issues of multiple testing in the four sequential stages that we used, we evaluated by simulation the likelihood of achieving the observed results by chance. All four stages of analysis and three rounds of marker selection were incorporated into the simulation (see Methods), and we obtained an empirical *P* value of 0.005, indicating that identification of two markers in stage 4 with the observed *P* values is highly unlikely to be a chance finding. Combining overall genotype data for stages 1–4 gave an OR of 1.19 ($P = 6.40 \times 10^{-9}$) for the 8q24

Table 1 SNPs associated with colorectal cancer disease status in the stage 3 and 4 study sets

ARCTIC SNP	Stage 3 tag SNP (if used)	Scotland stage 3 *P* value	Scotland stage 4 *P* value	Odds ratio (s.e.m.)
rs719725	rs7857628, rs206636213	0.0025	0.037	1.10 (0.05)
rs10483802	rs910315	0.00317	0.705	1.05 (0.11)
rs10489565	Direct genotype	0.0258	0.967	0.99 (0.07)
rs10505477	rs6983267	0.0305	0.001	1.16 (0.05)
rs10516168	Direct genotype	0.0484	0.105	1.12 (0.07)
rs11236164	Direct genotype	0.0506	0.168	0.94 (0.05)
rs10493889	Direct genotype	0.0508	0.814	0.98 (0.09)
rs850470	Direct genotype	0.0661	0.761	1.02 (0.05)
rs10491268	Direct genotype	0.0736	0.44	1.04 (0.05)

For each marker from the Ontario data set, the relevant tag SNPs are given, along with the *P* values for allelic tests of association in stages 3 and 4. Stage 4 allelic odds ratios and standard error (s.e.m.) are also shown.

Table 2 Genotyping results using the TaqMan assay at rs10505477 and rs719725 at each stage of this study

	Affected individuals			Controls						
rs10505477	GG	AG	AA	GG	AG	AA	OR	95% c.i.	*P*	P_{homo}
Ontario	249	557	361	305	586	297	1.22	1.09–1.37		
Newfoundland[a]	103	213	129	85	176	105	1.01	0.83–1.22		
Seattle[a]	141	333	218	177	327	190	1.20	1.03–1.38		
Newfoundland/Seattle							1.11	0.94–1.31		
Scotland-3	179	440	280	230	456	241	1.22	1.07–1.39		
Scotland-4	384	970	556	483	988	514	1.16	1.06–1.27		
France-Nantes	248	504	293	294	551	272	1.13	1.00–1.27		
France-Familial	76	192	102	136	291	112	1.28	1.05–1.55		
EPIC	167	372	222	189	365	195	1.13	0.98–1.31		
Stages 1–4							1.18	1.12–1.25	1.41×10^{-8}	0.74
French/EPIC							1.16	1.07–1.26	5.05×10^{-4}	0.54
All cohorts 1–7							1.17	1.12–1.23	3.16×10^{-11}	0.85
rs719725	CC	AC	AA	CC	AC	AA	OR	95% c.i.	*P*	P_{homo}
Ontario	138	502	508	159	581	439	1.20	1.07–1.36		
Newfoundland	66	208	162	64	156	142	1.01	0.83–1.23		
Seattle[a]	83	324	278	101	337	253	1.15	0.99–1.35		
Newfoundland/Seattle							1.09	0.96–1.24		
Scotland-3	117	410	353	139	447	314	1.17	1.02–1.34		
Scotland-4	264	895	753	301	955	713	1.10	1.01–1.21		
France-Nantes	165	510	363	175	537	389	0.99	0.88–1.12		
France-Familial	60	175	144	78	249	220	0.92	0.77–1.12		
EPIC	121	354	289	108	379	279	0.99	0.86–1.15		
Stages 1–4									1.32×10^{-5}	0.61
French/EPIC									0.61	0.80
All cohorts 1–7									0.023	0.11

In addition to the genotype counts, odds ratios (OR) based on a log-additive model are shown with their 95% confidence intervals (95% c.i.). Random effect model summary odds ratios, confidence intervals and *P* values are also shown, along with *P* values for tests for homogeneity (P_{homo}). Results from Newfoundland and Seattle were combined using a random effect model, and the combination was considered to be the stage 2 result.

locus (rs10505477 or rs6983267) and 1.13 ($P = 4.98 \times 10^{-5}$) for the 9p24 locus (rs719725 and rs7857826).

Communication among a number of groups investigating the genetics of colorectal cancer led to sharing of data at the 8q24 and 9p24 loci. A consortium genotyped 20 of 76 SNPs from stages 1 and 2 (including rs10505477 from 8q24) in 2,199 affected individuals and 2,401 controls derived from the European Prospective Investigation into Cancer and Nutrition (EPIC), Nantes and French Familial Case Control studies, resulting in validation of the association of rs10505477 with colorectal cancer (log-additive OR = 1.16; $P = 5.05 \times 10^{-4}$). None of the other 19 SNPs replicated any of the putative associations from stages 1 and 2 (**Supplementary Table 3** online). rs6983267 was also significantly different among affected individuals and controls in this same group (log-additive OR = 1.16, $P = 5.4 \times 10^{-4}$). The choice of the 20 loci genotyped in these cohorts was independent of the choice of the nine loci genotyped in our stage 4. Hence, we do not expect a selection bias in the estimates of the OR from cohorts derived from France and Europe as may be true for our results from stages 1–4.

To ensure technical consistency across all study sets, we genotyped rs10505477 (8q24) and rs719725 (9p24) using TaqMan assays in all 7,480 affected individuals and 7,779 controls. Genotype concordance between Affymetrix and TaqMan was high (3,865/3,880 individuals, or 99.61%). As expected, rs10505477 showed association with colorectal cancer in each set of subjects and in a pooled data set (OR = 1.17, 95% c.i. 1.12–1.23, $P = 3.16 \times 10^{-11}$) (**Table 2** and **Fig. 2**), and the effect seems to be additive or log additive with each copy of allele 'A'. Although the combined stages 1–4 provided evidence for an association with colorectal cancer at the 9p24 locus (OR = 1.14, $P = 1.32 \times 10^{-5}$), the association of rs719725 was not replicated in the EPIC and French data sets individually. When combined with stages 1–4, the result was only marginally significant (OR = 1.08; $P = 0.023$) (**Table 2** and **Fig. 2**). Despite this apparent lack of replication, it remains possible that this marker is a susceptibility locus, but further replication will be required in additional case-control sets.

Markers rs10505477 and rs6983267 are located on chromosome 8q24 within a 17,196-bp region of high linkage disequilibrium (LD) (**Fig. 3**). CEU HapMap estimates of LD between rs10505477 and rs6983267 ($r^2 = 0.94$) seem robust, as $r^2 = 0.95$ and $D' = 0.98$ in 325 Scottish subjects genotyped for both SNPs, indicating that both SNPs efficiently tag each other. Indeed, there are effectively only two common haplotypes (with frequencies of 0.552 and 0.435) because there are very few recombinations in the 5,862 bp between rs10505477 and rs6983267.

Chromosome 8q24 polymorphisms have recently been reported to be associated with prostate cancer risk. At least three risk-associated regions, separated by sites of recombination, confer independent risk[1–3,9]. The haplotype block containing SNP rs6983267, which we

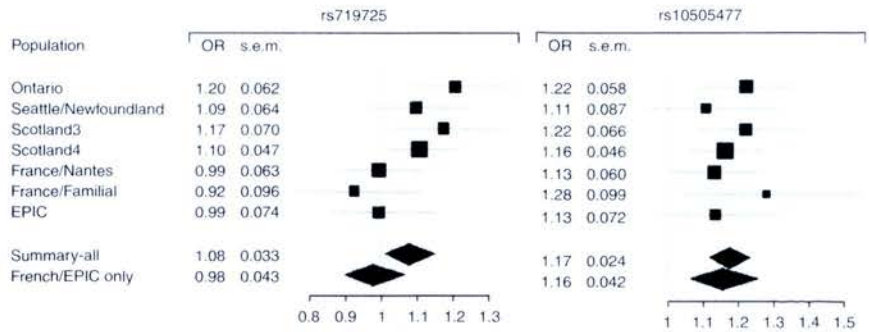


Figure 2 Odds ratios associated with an additional copy of allele A at rs1050477 or allele A at rs719725 in samples of affected individuals and controls from our four-stage analysis and validation in several populations, assuming a log-additive impact on disease risk. Ontario refers to affected individuals and controls from the OFCCR. Seattle/Newfoundland is the combined result from the Seattle and Newfoundland data sets. Scotland 3 refers to early-onset affected individual and a matched control set from Scotland. Scotland 4 refers to later-onset affected individuals and matched controls. France/Nantes refers to affected individuals ($n = 1068$) and controls ($n = 1113$) from the Nantes cohort, France/Familial refers to affected individuals ($n = 370$) and controls ($n = 539$) from the Familial Cohort. EPIC refers to affected individuals ($n = 761$) and controls ($n = 749$) from the European Prospective Investigation into Cancer and Nutrition. French/EPIC refers only to the combination of the Nantes, Familial and EPIC study sets. Symbols are plotted proportionally in size to the number of affected individuals in each data set. Random effect models were used to combine results across populations[8]. Lines to the left and right of symbols represent 95% confidence intervals.

identified as conferring risk of colon cancer, also confers prostate cancer risk with an OR of 1.58 for subjects homozygous for the risk allele[3].

The colon and prostate cancer risk locus at 8q24 contains two ORFs. One, DQ515897, is an uncharacterized gene with multiple alternatively spliced products, and rs10505477 lies within intron 6. The biological relevance of this alternatively spliced gene product is unknown. The second ORF, DQ486513, is located 20,414 bp telomeric to rs10505477 and is currently categorized as a putative pseudogene of the transcription factor *POU5FP1*. There are several regions of sequence homology with DQ486513 in the human genome, including those located on chromosomes 1, 6, 10 and 12. However, a 50-bp sequence centered on rs10505477 appears unique to the chromosome 8q locus.

In addition to its close proximity to DQ515897 and DQ486513, rs10505477 is also 340,873 bp telomeric to the oncogene *MYC*, which has a known role in colon cancer biology. We performed immunohistochemistry on formalin-fixed paraffin embedded tumor tissue from genotyped individuals ($n = 86$) from the OFCCR to assess MYC protein expression in affected individuals. However, there was not a statistically significant relationship between genotype and MYC expression (**Supplementary Table 4** online).

The Canadian and French groups initiated their projects with a hypothesis-driven candidate gene approach. They selected a common set of 1,536 SNPs using an early version of the HapMap for 227 candidate genes, of which 138 are implicated in DNA repair and the remainder in apoptosis, chromosome stability, cell cycle control or nucleotide metabolism. The lack of significant hits in an early joint analysis led to a change in strategy toward a genome-wide approach. Although 330 SNPs from this set were investigated in the Newfoundland and Seattle samples, and a subset of these moved to later stages, the two loci that were subsequently replicated in Scotland were from the genome-wide SNP panels and not from the hypothesis-based list of genes. Using our four-stage strategy, the power in this study was excellent for detecting a locus with the observed magnitude of effect, assuming that at least five true loci of such magnitude were in strong LD with our markers. However, we acknowledge that the strategy used here may miss variants that influence colorectal cancer risk, if such variants are rare, if they exert very small effects, if they interact with environmental factors or other genes or if our markers are not in sufficiently strong LD with such risk variants. Further studies, using higher-density SNP arrays to increase genome coverage, are ongoing, which may lead to the identification of additional common risk alleles for colorectal cancer.

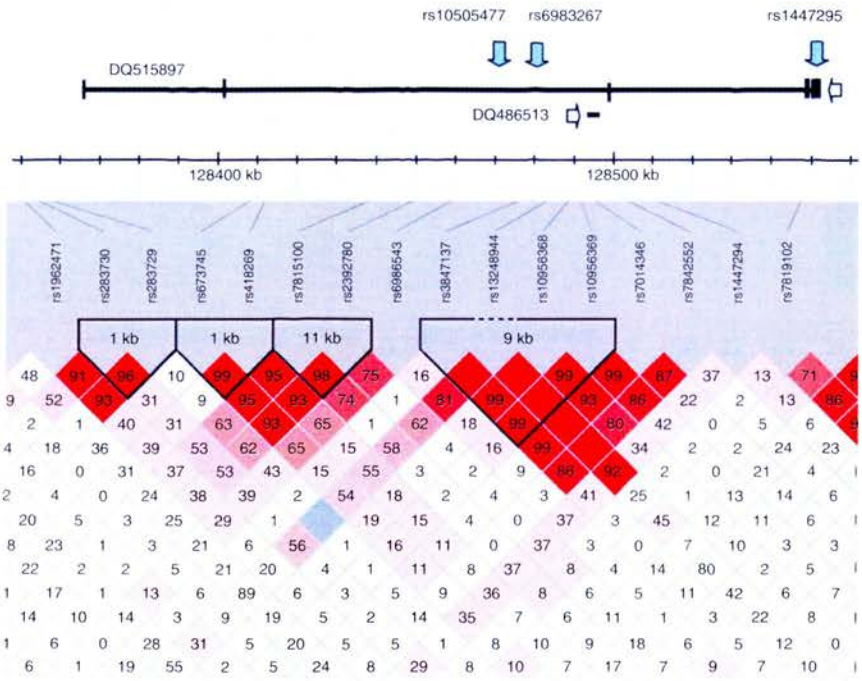


Figure 3 LD and locus map of 8q24. LD coefficients (r^2) of stage 1 SNP are indicated. The position and transcription orientation of DQ515897 and DQ486513 mRNA are indicated. SNPs rs10505477 and rs6983267, found within the indicated region of LD, are associated with the colon cancer phenotype.

This study shows definitively that a locus on chromosome 8q close to SNPs rs10505477 and rs6983267, within a small region of tight LD, is associated with colorectal cancer susceptibility. This is common genetic variant was identified by a genome-wide scan using genetic association to successfully identify and validate a genetic risk factor for large bowel malignancy. Although it is unlikely that this finding in itself will have direct clinical relevance for the individual, this discovery will lead to better understanding of colorectal cancer biology and disease causation. Furthermore, this marker, in conjunction with others yet to be discovered, could be used in a population setting to stratify risk. Groups with marginally elevated risk could be identified so that they might benefit from a tailored approach to screening and/or preventative interventions.

METHODS

Study populations. Individuals with colorectal cancer and population controls participating in three familial cancer registries (two in Canada, one in the US) were used in the initial phases of this study. Cases and controls from the Ontario Familial Colorectal Cancer Registry (OFCCR) were used for the initial analyses that detected associations, and then the Newfoundland Familial Colorectal Cancer Registry (NFCCR) and the Seattle Familial Colorectal Cancer Registry (SFCCR) were used to assess consistency in different populations. Stage 3 and 4 validation studies were performed on samples from the Scottish Colorectal Cancer Genetic Susceptibility Study (COGS) cohort. Further validation was performed using the Nantes, the EPICS and the French Familial Cohort. Details of these populations are provided in **Supplementary Methods**. All subjects provided written informed consent. This study was approved by the ethics review boards of the Toronto Academic Health Sciences Council, the Fred Hutchinson Cancer Research Center and Memorial University (Newfoundland).

SNP panels used by study stage. Stage 1 genotyping used three separated SNP panels. The first included 1,502 SNPs from 227 genes, and 34 random SNPs (http://fisher.utstat.toronto.edu/~celia/index.php) anticipated to be of relevance to colorectal cancer biology (**Supplementary Table 5** online). Panel 2 SNPs involved the Affymetrix GeneChip Human Panel 1 10K array, that included nonsynonymous double-validated public cSNPs, from approximately 8,000 genes[10]. The complete SNP and gene list can be found at http://www.affymetrix.com/products/reagents/specific/application_specific.affx.

Panel 3 SNPs on the Affymetrix GeneChip Human Mapping 100K set were selected by the manufacturer (http://www.affymetrix.com/support/technical/manual/taf_manual.affx), producing an average distance between markers of 26 kb[11].

In stage 2, when selecting markers for the first validation, we allowed our significance thresholds to vary by platform. We expected differing rates of true positive associations to occur across the three data sets used in stage 1, owing to marker selection. For example, we expect more true positive associations in the Illumina custom set (selected at candidate gene loci) than in the Affymetrix 100K set. We identified the optimal configuration of markers to minimize the expected false discovery rate[7]. Markers (1,152) were selected for validation, of which 1,143 were successfully genotyped: 330 from the Illumina GoldenGate platform, 142 from the Affymetrix 10K cSNP platform and 671 from the 100K Affymetrix GeneChip. In stage 3, genotypes (76) were obtained from a concurrent genome-wide scan and by direct genotyping. The early onset case-control set were genotyped using Illumina HumanHap300 and HumanHap240S arrays genotyped on the Infinium platform and by TaqMan assay for sites not represented on these panels. In stage 4, genotypes (9) in the Scottish cohort, which was unrestricted by age, were measured by TaqMan assay. Selection of SNPs and genotyping of the Nantes, EPICs and French Familial study were based on the Illumina HumanHap300 and HumanHap240S arrays.

Statistical analysis. To obtain the Ontario data, using 10,000 permutations of case-control labels, disease-marker associations were calculated for each marker by calculating the empirical significance associated with the largest of three test statistics: dominant, recessive and a test for trend[12]. False discovery rates were estimated[13]. Markers were selected for replication in the Seattle and Newfoundland datasets by minimizing the platform-stratified false discovery rate[8] (see SNP panels, Stage 2, above). This led to different *P*-value selection thresholds for the three genotyping datasets on the Ontario data.

For the first replication, data from two sources, Seattle (698 affected individuals, 700 controls) and Newfoundland (452 affected individuals, 367 controls), were combined. For each marker and each population, dominant, recessive and trend log odds ratios and empirical estimates of each s.e.m. were obtained, the latter by dividing the log odds ratio by the normal statistic associated with the empirical *P* value obtained from 10,000 permutations. Stage 2 disease-marker associations were estimated by combining the two population-specific log odds ratios using random effect models[8]. The *P* value for association was obtained by testing whether the summary log odds ratio was zero.

Model-specific odds ratios were compared between stage 1 and stage 2 to determine direction of effect. To obtain an odds ratio measure corresponding to the trend test, we calculated the log-additive odds ratio corresponding to a linear coding of the number of high-risk alleles in a logistic model.

For stages 3 and 4, tests of disease-marker associations were calculated using allelic models. The probability that two markers remain significant after three consecutive rounds of (i) selection of a subset of significant markers and (ii) testing in new independent data was evaluated via a simulation study. Specifically, at each stage j of the selection process, we generated N_j *P*-values and N_j 'directions', where N_j is the number of markers genotyped in stage j, $N_1 = 99{,}632$, $N_2 = 1{,}143$, $N_3 = 9$ and $N_4 = 2$. The *P* values were generated from the uniform distribution, assuming that no markers were associated with disease, and the 'directions', representing which allele was the higher risk allele, were generated from Bin(0.5). We sorted the *P* values and selected the $N_j + 1$ smallest *P* values with a matching direction of effect. After selecting $N_4 = 2$, we counted the number of simulations where the first three *P* values for these two markers were <0.10, the last (in stage 4) was less than 0.05 and all four directions matched. This empirical significance was $P = 0.005$, implying that these two markers are unlikely to be significant by chance alone.

We calculated power assuming there were one, five or ten loci in strong LD with one of our markers, with the same differences in allele frequency as observed at rs10505477 (49% versus 55%). Our power simulations used a similar design as the type 1 error calculations described in the previous paragraph, framed around our four-stage analytic strategy, including the selection at each stage, but including these fixed numbers of true positives. For only one true positive locus, the most significant marker is likely to be the true locus only 43% of the time. For five true positives, in 90% of simulations, the most significant marker will be a true locus, and for ten true positives, the estimated power is 100% (out of 150 simulations).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors thank D. Daftary and T. Selander at the Mount Sinai Hospital Biorepository for technical and administrative assistance with OFCCR samples. We also are grateful to A. Bélisle, S. Roland, M.-C. Tessier and D. Vincent of the McGill University and Génome Québec Innovation Centre for technical genotyping assistance. We acknowledge all those involved in recruitment and assembly of the biological and data resources of the Colorectal Cancer Genetic Susceptibility (COGS) study and the Scottish Colorectal Cancer Study (SOCCS), including the Edinburgh Wellcome Trust Clinical Research Facility and also the Family Practitioner Services Department, the Cancer Intelligence Unit of the Information and Statistics Division (ISD) and Scottish Cancer Registry Cancer, all of the Scottish Central National Health Service (NHS). C. Bonithon-Kopp, A. Pariente, B. Nalet and J. Lafon (Group d'Etude des Adenomas) and members of the Association Nationale des Gastroenterologues des Hopitaux Généraux. For administrative assistance, we acknowledge L. Blahut (Cancer Care Ontario). We are grateful to the nursing, laboratory and office staff throughout Edinburgh, at the Wellcome Trust Clinical Research Facility and at the central Scottish NHS departments, including Cancer Registry and the Scottish Cancer Intelligence Unit of ISD. Cancer Care Ontario, as the host organization to the ARCTIC Genome Project, acknowledges that this Project was funded by Genome Canada through the Ontario Genomics Institute, by Génome Québec, the Ministère du Dévelopement Économique et Régional et de la Recherche du Québec and the Ontario Institute for Cancer Research (B.W.Z., T.J.H., C.M.T.G, S.G. and M.C.). This work was supported through collaboration and cooperative agreements with

the Colon Cancer Family Registry and principal investigators supported by the US National Cancer Institute, US National Institutes of Health under RFA CA-95-011, including S.G. at the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA076783) and J.P. at the Seattle Colorectal Cancer Family Registry (U01 CA074794). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating institutions or investigators in the Colon Familial Registry, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government or the Colon CFR. This work was supported through collaboration and cooperative agreement with the Newfoundland Familial Colon Cancer Registry at the Memorial University of Newfoundland (B.Y., R.G., J.G.). The work in Scotland was supported by Cancer Research UK (C348/A3758, C48/A6361), the UK Medical Research Council (G0000657-53203) and the Scottish Executive Chief Scientist's Office (K/OPR/2/2/D333, CSO CZB/4/94) (M.D.) and by a Centre Grant from the Digestive Disorders Foundation (http://www.corecharity.org.uk). Support in France came from the French Ministry of Research, Fondation de France (S.O., C.B.P.), Projet Hospitalier de Researche Clinique (PHRC) AOM01-006 (G.T.), Ligue Nationale contre le Cancer (G.T.), Groupement des Enterprises Francaises dans la Lutte contre le Cancer (GEFLUC) (S.B.) and the European Commission (E.R.). Support is acknowledged from the National Program for Complex Data Structures (Canada) (R.K.) and the Centre for Applied Genomics (Toronto) (C.G.).

AUTHOR CONTRIBUTIONS

B.W.Z., C.M.T.G., R.K., S.G., M.C. and T.J.H. devised and executed the original study design and prepared this manuscript. J.R., A.T., S.M.F., J.P., S.O., T.C., E.C., V.F., P.L., S.S., S.R., J.-F.O., F.R. and A.M. assisted in data analysis. R.S., P.C., S.B., A.M.O., G.Z., M.J., J.M., B.Y., R.G., J.G., M.E.M.P., H.C., H.B., M.S., E.T., C.B.-P., B.B., E.R., S.K. and S.J.C. provided insights and G.T., M.G.D., J.P., P.N. and E.P. assisted in sample procurement and study design and interpretation. A.T., S.M.F., J. Prendergast, M.E.M.P., H.C. and M.G.D. designed and undertook stages 3 and 4 of the study using the Scottish sample and data resource and contributed to writing the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
2. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
3. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
4. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
5. Cotterchio, M. *et al.* Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis. Can.* **21**, 81–86 (2000).
6. Fan, J.B. *et al.* Illumina universal bead arrays. *Methods Enzymol.* **410**, 57–73 (2006).
7. Greenwood, C.M.T., Rangrej, J. & Sun, L. Optimal selection of markers for validation from genome-wide association studies. *Genet. Epidemiol.* **31**, 396–407 (2007).
8. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
9. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
10. Wang, Y. *et al.* Allele quantification using molecular inversion probes (MIP). *Nucleic Acids Res.* **33**, e183 (2005).
11. Nicolae, D.L., Wen, X., Voight, B.F. & Cox, N.J. Coverage and characteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. *PLoS Genet.* **2**, e67 (2006).
12. Sasieni, P.D. From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261 (1997).
13. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* **59**, 289–300 (1995).

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Chromatin structure and evolution in the human genome

James GD Prendergast*[1], Harry Campbell[3], Nick Gilbert[2], Malcolm G Dunlop[1], Wendy A Bickmore[2] and Colin AM Semple[2]

Address: [1]Colon Cancer Genetics Group, Division of Oncology, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK, [2]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU,UK and [3]Public Health Sciences, Department of Community Health Sciences, University of Edinburgh, Edinburgh, UK

Email: James GD Prendergast* - James.Prendergast@hgu.mrc.ac.uk; Harry Campbell - Harry.Campbell@hgu.mrc.ac.uk; Nick Gilbert - Nick.Gilbert@ed.ac.uk; Malcolm G Dunlop - Malcolm.Dunlop@hgu.mrc.ac.uk; Wendy A Bickmore - Wendy.Bickmore@hgu.mrc.ac.uk; Colin AM Semple - Colin.Semple@hgu.mrc.ac.uk

* Corresponding author

Published: 9 May 2007

BMC Evolutionary Biology 2007, **7**:72 doi:10.1186/1471-2148-7-72

Received: 16 November 2006
Accepted: 9 May 2007

This article is available from: http://www.biomedcentral.com/1471-2148/7/72

Abstract

Background: Evolutionary rates are not constant across the human genome but genes in close proximity have been shown to experience similar levels of divergence and selection. The higher-order organisation of chromosomes has often been invoked to explain such phenomena but previously there has been insufficient data on chromosome structure to investigate this rigorously. Using the results of a recent genome-wide analysis of open and closed human chromatin structures we have investigated the global association between divergence, selection and chromatin structure for the first time.

Results: In this study we have shown that, paradoxically, synonymous site divergence (dS) at non-CpG sites is highest in regions of open chromatin, primarily as a result of an increased number of transitions, while the rates of other traditional measures of mutation (intergenic, intronic and ancient repeat divergence as well as SNP density) are highest in closed regions of the genome. Analysis of human-chimpanzee divergence across intron-exon boundaries indicates that although genes in relatively open chromatin generally display little selection at their synonymous sites, those in closed regions show markedly lower divergence at their fourfold degenerate sites than in neighbouring introns and intergenic regions. Exclusion of known Exonic Splice Enhancer hexamers has little affect on the divergence observed at fourfold degenerate sites across chromatin categories; however, we show that closed chromatin is enriched with certain classes of ncRNA genes whose RNA secondary structure may be particularly important.

Conclusion: We conclude that, overall, non-CpG mutation rates are lowest in open regions of the genome and that regions of the genome with a closed chromatin structure have the highest background mutation rate. This might reflect lower rates of DNA damage or enhanced DNA repair processes in regions of open chromatin. Our results also indicate that dS is a poor measure of mutation rates, particularly when used in closed regions of the genome, as genes in closed regions generally display relatively strong levels of selection at their synonymous sites.

Background

Regions of open and closed chromatin structure have recently been defined across the human genome [1]. Gilbert et al showed that regions of open chromatin are often gene dense and appear to correlate well with clusters of broadly expressed genes. They suggested that open chromatin fibre domains provide a chromatin environment more conducive to transcriptional activation. However, many genes are also found in regions of closed chromatin structure. This raised the question as to why would genes be maintained in closed chromatin if this meant they were simply less accessible for transcription. One possibility is that they need to be subject to especially tight transcriptional regulation, and that their aberrant or leaky expression in inappropriate cells cannot be tolerated. However, it has also been proposed that open chromatin structure may make the underlying DNA sequence more susceptible to DNA damage [2].

Although some studies have predicted that rates of mutation are relatively constant across mammalian genomes, analysis of human-mouse alignments has suggested that there may be as much as a 3-fold difference in substitution rates across chromosomes [3], with regions containing genes involved in extracellular communication displaying unusually high levels of synonymous substitutions [4]. Previous studies have also shown that, in mammals, genes within close genomic proximity undergo similar rates of neutral divergence and evolution [4-6]. For example, Williams and Hurst showed that the mean difference between the Ka values (substitution rate at non-synonymous sites) of 176 pairs of linked genes was significantly lower than would be expected by chance [5]. Similar results were also observed with Ks (substitution rate at synonymous sites) and Ka/Ks (often used to infer the mode and strength of selection). Consequently they proposed that the murid genome was split into domains of evolution. The reason for this was unknown, but it is possible that some aspect of chromatin structure over different genomic regions influences the rate of DNA damage or its repair.

The availability of a map of long-range chromatin structure across the human genome [1] allows us to assess this idea and, through the comparison of various measures of neutral variation, we have identified those forms of chromatin structure associated with the highest rates of background mutation.

Results and discussion

Non-dS measures of mutation are highest in closed chromatin

In order to determine whether background mutation rates are associated with chromatin structure we first determined intergenic divergence rates, using human versus chimpanzee whole genome alignments, in regions whose chromatin environment in human lymphoblastoid cells had been determined. The majority of intergenic bases should be under little or no selection and therefore intergenic divergence should be approximately analogous to background mutation rates. As shown in Figure 1A, we found a negative correlation between intergenic divergence and chromatin structure at non-CpG sites. As open chromatin is generally more gene rich than closed (and may therefore contain more regulatory elements than intergenic regions) we also examined divergence rates in ancient repeats only. However, these also displayed the lowest divergence rates when in open chromatin (Figure 1B).

It has previously been proposed that DNA sequences nearer the centre of the nucleus may be protected from DNA damage by those on the periphery (the "bodyguard hypothesis"). Likewise, the chromosomes most enriched with open chromatin are generally situated towards the centre of a nucleus [2]. The correlation observed between divergence rates and chromatin structure may therefore be an indirect result of these phenomena. We therefore investigated whether a correlation between intergenic divergence and chromatin structure could be observed within chromosomes. Although chromosomes themselves have been shown to display some level of polar organization (such that their most gene-poor regions are those closest to the nuclear periphery [7]) adjacent intergenic regions within chromosomes often have very different chromatin structures despite displaying approximately the same nuclear localisation. If the observed correlation between intergenic divergence and chromatin structure reflects the predictions of the bodyguard hypothesis we would expect to see no such correlation within chromosomes. This, however, is not the case. For example, as shown in Figure 1C, there is a significant negative correlation between intergenic divergence and chromatin structure within chromosome 1 ($r^2 = 0.053$; $p = 0.0043$). The two outlier clones observed in this figure, with a divergence greater than 0.025, could represent mutational hotspots in the genome. However, the degree of difference between the divergence observed in these clones compared with the rest of the chromosome suggests to us that the alignments in these regions are more likely be of poor quality. Removal of these clones increases the significance of the correlation observed between divergence and chromatin structure ($r2 = 0.113$; $p = 2.5e\text{-}05$). In total 7 out of 22 chromosomes display a significant negative correlation ($p < 0.05$) between clone intergenic divergence and chromatin structure (Chromosomes 1, 2, 5, 8, 12, 17 and 20). These data therefore argue against the bodyguard hypothesis being solely responsible for these observed correlations between chromatin structure and intergenic divergence rates.

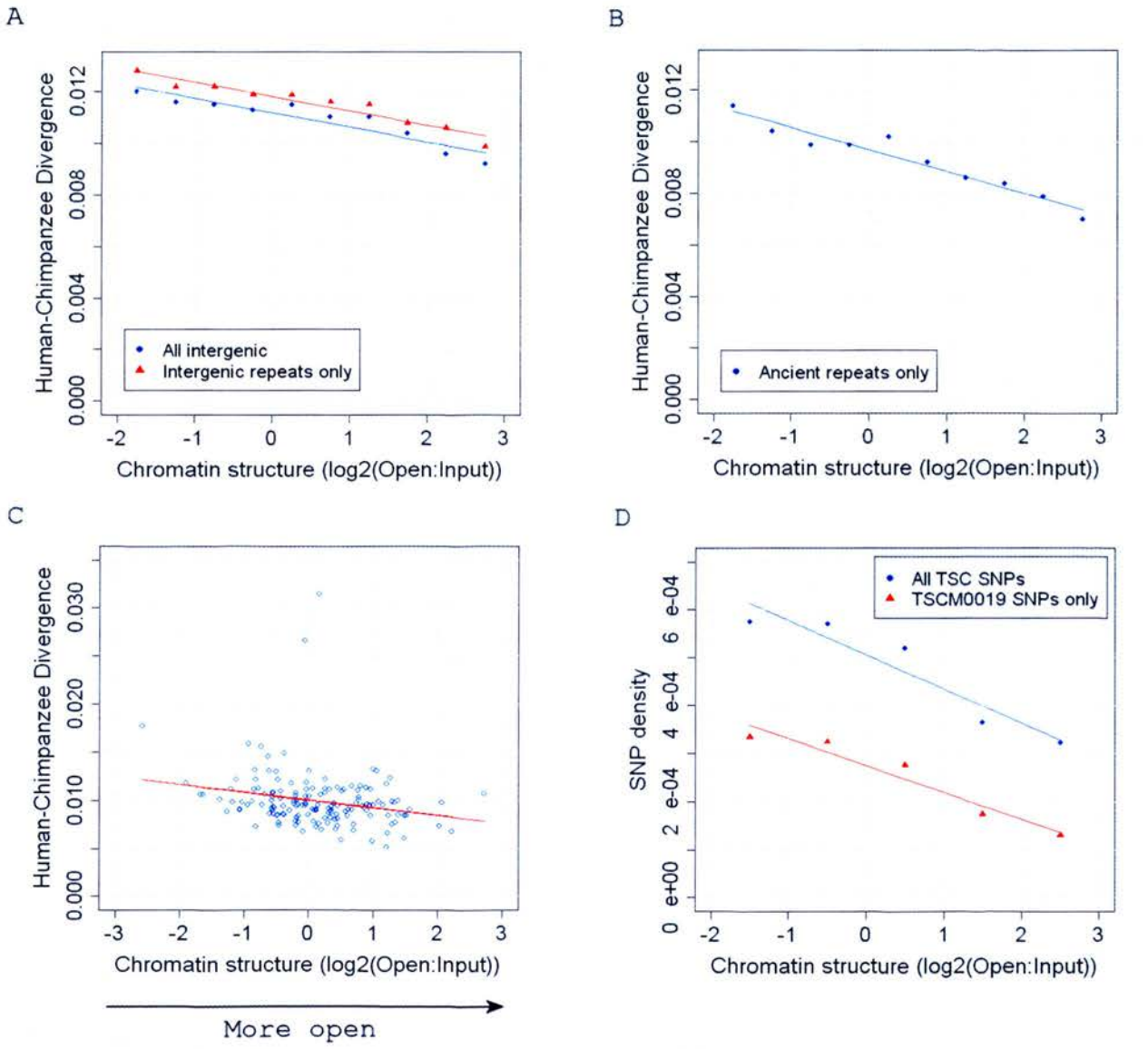


Figure 1

Increased mutation rates in closed chromatin at non-CpG sites. (A+B) Mean intergenic and ancient repeat divergence observed across chromatin categories (Intergenic r^2: = 0.87, p = 9.1e-05; Intergenic repeats only r^2: = 0.93, p = 7.3e-06; Ancient repeats only r^2: = 0.93, p = 6.1e-06). (C) Intergenic divergence of each 1 Mb clone from chromosome 1 against their corresponding chromatin score (10 clones containing less than 10,000 intergenic bases were excluded). (D) Mean human SNP densities (SNPs/bp) observed across chromatin categories (All SNPs r^2: = 0.89, p = 0.016; Single random detection protocol (TSCM0019) SNPs only r^2: = 0.93, p = 0.008).

Another measure often used to predict mutation rate is SNP density [8,9]. It is predicted that as a large proportion of intergenic sequence is non-functional and that there has been little time for selection to act on SNPs, their density along the genome should generally reflect underlying mutation rates. A further benefit of the use of SNPs in this way is that mutation rates can be predicted without relying on sequence comparisons with other species. We consequently determined the mean intergenic SNP densities observed across chromatin categories. As shown in Figure 1D the mean SNP density was also lowest in the most open regions of the genome.

There is therefore strong evidence that mutation rates are associated with chromatin structure. Not only are intergenic, intronic (Figure 2A) and ancient repeat divergence rates highest in closed chromatin but the density of SNPs is also elevated in the most closed regions of the human genome. Thus we hypothesise that closed regions of the genome are simply less accessible to DNA repair mechanisms.

It should be noted however that chromatin structure is likely to be only one of several factors associated with neutral divergence rates in the human genome. This is most apparent on chromosome 19, and to a lesser extent chromosome 8, which show substantially higher mean intergenic divergence rates in our analysis than the other autosomes. Whereas chromosome 19 and chromosome 8 display mean intergenic divergences of 1.5% and 1.3% respectively, the divergence rates of all other autosomes fall between 1 and 1.2%. As chromosome 19 is particularly enriched with open chromatin [1], its high divergence levels are contrary to what is generally observed across the autosomes. The high levels of divergence observed along chromosome 19 are consequently likely to be a result of factors other than chromatin structure.

Gene distribution and chromatin structure

As shown in figure 3A, housekeeping genes are generally located in the more open regions of the genome and tissue-specific genes in the most closed regions. This is in agreement with a previous analysis that illustrated that nucleosome formation potential is negatively correlated with expression breadth [10]. Consequently a recent study by Gazave et al. [11], that showed that the levels of human-chimpanzee divergence observed in the introns of housekeeping genes is significantly lower than in other genes, is in broad agreement with the analysis presented here. Although CpG dinucleotides were not excluded in Gazave et al's analysis this is only likely to have led to an increase in the estimation of divergence in housekeeping genes due to the enrichment of CpG dinucleotides in open chromatin. However, intriguingly, when human-mouse alignments are examined, the introns of tissue-specific genes have been shown to contain a greater proportion of conserved sequence than those of housekeeping genes [12] (in contradiction to what is observed in human-chimpanzee alignments). We believe this apparent discrepancy is likely to be the result of the difference in evolutionary distance investigated, with the examination of human-mouse alignments potentially leading to the identification of regions under (stabilising) selection. For example, we may expect that closed regions of the genome contain more DNA elements involved in regulating the surrounding chromatin structure whose conservation becomes apparent across larger evolutionary distances.

Through the use of the DAVID program [13] that determines those biological terms and annotations (for example GO terms) enriched among a set of genes, we identified further classes of genes most over-represented in closed chromatin, and therefore likely to be experiencing the highest mutation rates. Of the 148 genes in the most closed regions of the genome, 40 encode glycoproteins (p for enrichment: 0.000074) and 22 were associated with the G-protein coupled protein signaling pathway ($p = 0.00011$). Glycoproteins and G-protein coupled receptors are involved in immune response and cell signaling and it has previously been proposed that such genes are likely to evolve quickly in response to changing stimuli [4]. Being located in closed regions of the genome (where we have observed background mutation rates (intergenic divergence and SNP density) are particularly high) will allow this more rapid evolution. Housekeeping genes, on the other hand, that are enriched in open chromatin, have previously been shown to evolve relatively slowly [14]. The location of a gene in the genome and its subsequent local chromatin structure may therefore, at least partly, be governed by the suitability of the local mutation rate it confers.

dS, unlike dN and dN/dS, is highest in regions of open chromatin

dS has historically been used as a further surrogate measure of basal mutation rates, as synonymous sites were believed to be under little or no selection. Changes at synonymous sites, unlike at non-synonymous sites, do not affect the encoded amino acid. In addition, due to the relatively small effective population sizes of mammals, a synonymous site would have to experience relatively strong selection to evolve in a non-neutral manner [15]. As shown in Figure 4A, the average rate of non-synonymous changes (dN) observed in human mouse alignments is 51% higher in the most closed chromatin regions of the genome than in the most open regions. Similarly, the ratio of non-synonymous to synonymous substitution rates (dN/dS), which is frequently used as a measure of selection, is 61% higher (Figure 4B). However, the average synonymous rate (dS) for genes in relatively open chromatin is higher than that for genes in a more closed chromatin structure (Figure 4C). This is consistent with the reported high Ks for human chromosome 19, the human chromosome with one of the most open chromatin structures of all [16]. The observation by Hurst et al. of similar levels of human-mouse dS, dN and dN/dS in linked genes is likely therefore to be the result of linked genes being from similar chromatin environments. To ensure the converse is not true, and that the results observed in this study are not the result of linked genes, we randomly selected only one gene from each clone (so that all genes analysed were approximately 1 Mb apart and therefore unlinked).

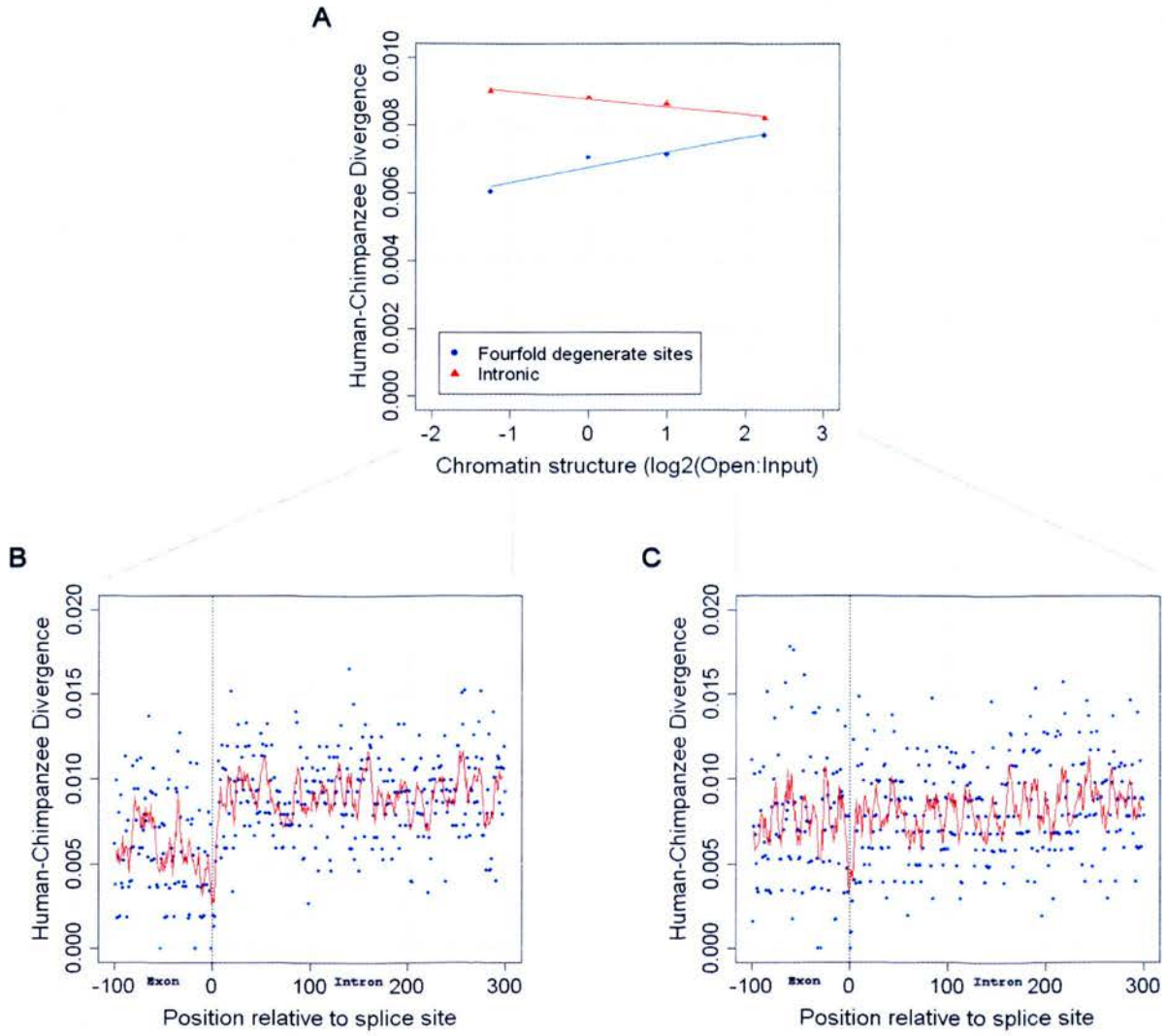


Figure 2

Comparison of splice site divergence observed across chromatin categories. (A) The divergence at non-CpG fourfold degenerate and intronic sites on autosomes only, with the divergence observed across the splice sites of the most closed (B) and open (C) genes shown below. (Closed exonic vs. closed intronic Mann-Whitney U test: p = 4.4e-16; open exonic vs. open intronic Mann-Whitney U-test: p = 0.053)

With this selection strategy we still observed similar correlations to those shown in Figure 3 (not shown).

Although we would expect the enrichment of housekeeping genes in relatively open regions of the genome as shown in Figure 3A (as open chromatin is likely to provide a more conducive environment for transcription), the lower average dN/dS observed in open chromatin may simply be a consequence of this higher number of housekeeping genes (which are known to evolve slowly) in these regions. The exclusion of housekeeping genes from the analysis, however, has little effect on the correlations in Figure 4 (not shown). Even the exclusion from the analysis of all genes whose 5' end is associated with a CpG island (which includes almost all housekeeping genes [17] and that are also enriched in open chromatin, Figure 3B) does not lead to the loss of the correlations between chromatin structure and dN, dS and dN/dS. In fact the rate of dN in CpG island genes, unlike that in genes not associated with a CpG island, is relatively constant across chromatin categories and does not show a significant correlation with chromatin. Consequently selection appears

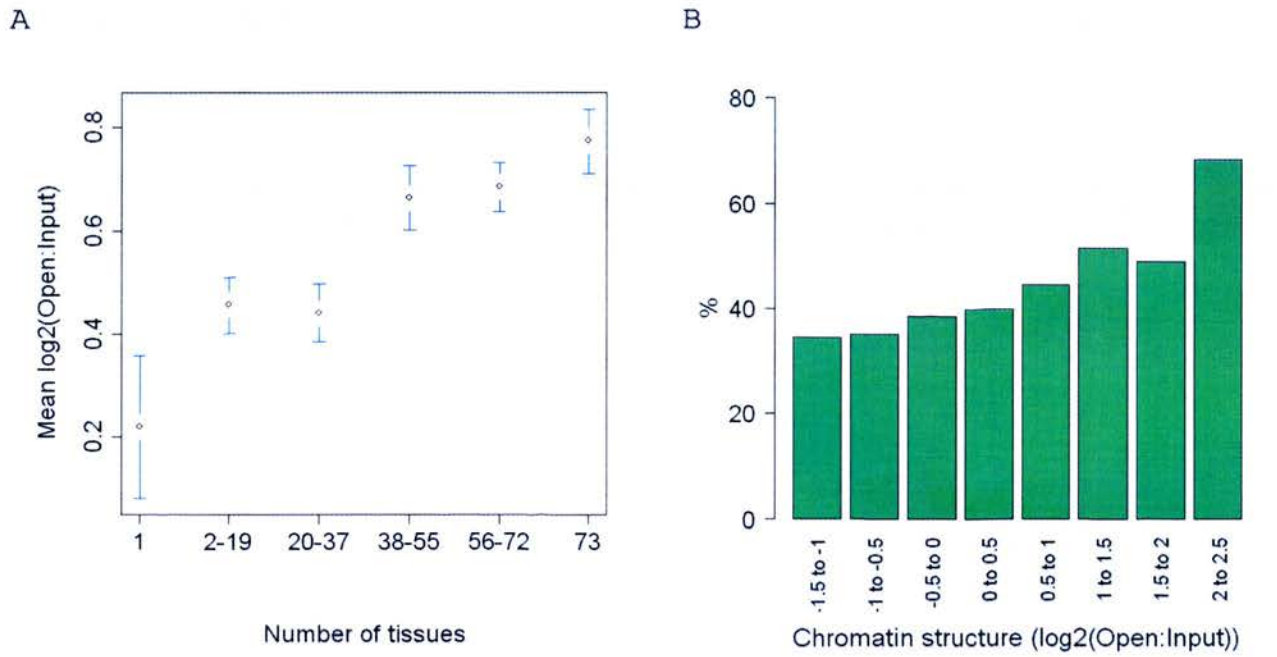


Figure 3
The distribution of gene expression profiles and CpG island genes across chromatin categories. (A) The mean chromatin structure (log2(Open:Input)) of genes of differing expression breadth across normal tissues (Kruskal-Wallis p = 7.5e-6) (B) The percentage of genes across chromatin categories that are associated with a CpG island.

to maintain similar levels of dN in genes associated with a CpG island irrespective of their local chromatin structure.

To ensure these results were not confounded by CpG associated or sex chromosome specific factors (sex chromosomes have been shown to display abnormal rates of divergence when compared to the autosomes [18]), we calculated divergence rates at non-CpG, fourfold degenerate sites in genes on autosomes only. We also used human-chimp alignments instead of human-mouse alignments as the chromatin structure of the chimp genome should be more similar to that in humans (and consequently the species of origin for each change is less important). However, as shown in Figure 2A, the highest rates of divergence are still observed in genes from the most open regions of the genome.

Genes in closed chromatin display the highest levels of selection at synonymous sites

Although historically the synonymous substitution rate (dS or Ks) has been used as a measure of the rate of mutation, there is increasing evidence that selection may be occurring at synonymous sites [15]. To investigate the potential role of any selection on synonymous sites in the disparity between dS and other measures of mutation, we analysed the rates of divergence observed across intron-exon boundaries [18]. As shown in Figure 2, the rates of intronic divergence in open regions of the genome are comparable to those observed at corresponding exonic, fourfold degenerate sites. This would suggest that genes in open chromatin display little if any evidence for selection at their synonymous bases. However, genes in closed chromatin display markedly higher rates of divergence at their intronic sites than at corresponding fourfold degenerate sites. Genes in closed chromatin therefore, unlike those in open, display strong evidence for synonymous site selection.

Although the rate of selection against both synonymous transitions and transversions is highest in closed chromatin, only the rate of synonymous transitions is strongly positively correlated with chromatin structure (Figure 5A). The rate of transversions at fourfold degenerate sites shows no obvious trend across chromatin categories (Figure 5B) and consequently selection against transversions, unlike transitions, appears to be independent of any factors associated with chromatin structure. We are not aware of any reason for the observed association between rates of transitions at non-CpG fourfold degenerate sites and

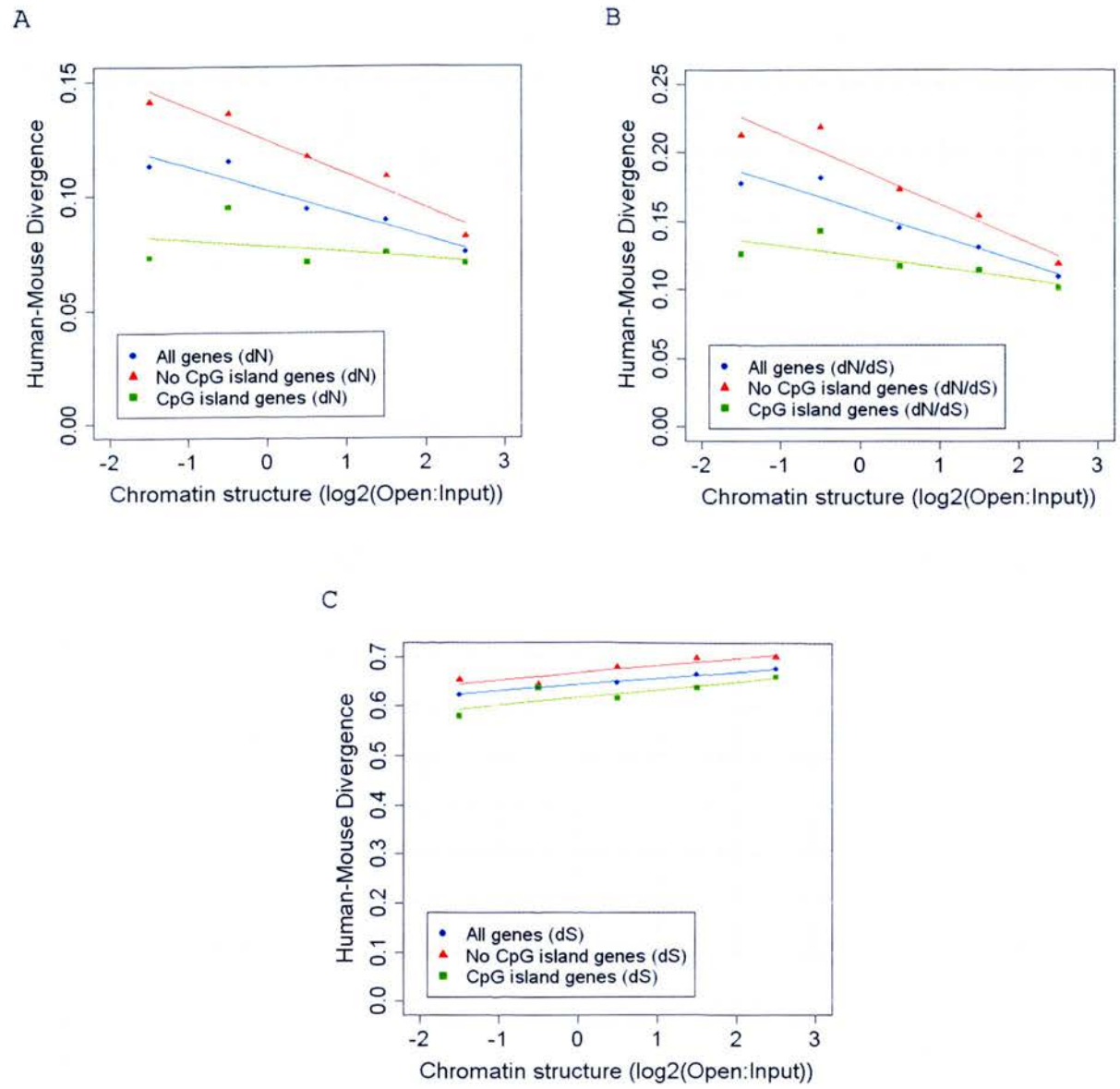


Figure 4

Human-mouse divergence across chromatin categories. Mean dN (A), dN/dS (B) and dS (C) in human/mouse coding sequence alignments. (All protein coding genes dS r^2: = 0.99, p = 0.001; dN r^2: = 0.92, p = 0.01; dN/dS r^2: = 0.92, p = 0.009. Genes associated with a CpG island dS r^2: = 0.72, p = 0.07; dN r^2: = 0.17, p = 0.5; dN/dS r^2: = 0.64, p = 0.1. Genes not associated with a CpG island only dS r^2: = 0.84, p = 0.03; dN r^2: = 0.95, p = 0.005; dN/dS r^2: = 0.92, p = 0.01.).

chromatin structure, but it could reflect constraint in motifs whose distribution are not uniform across the genome.

As previously shown, open regions of the genome are particularly gene dense whereas closed regions are relatively gene poor [1]. Consequently, the use of dS as a measure of

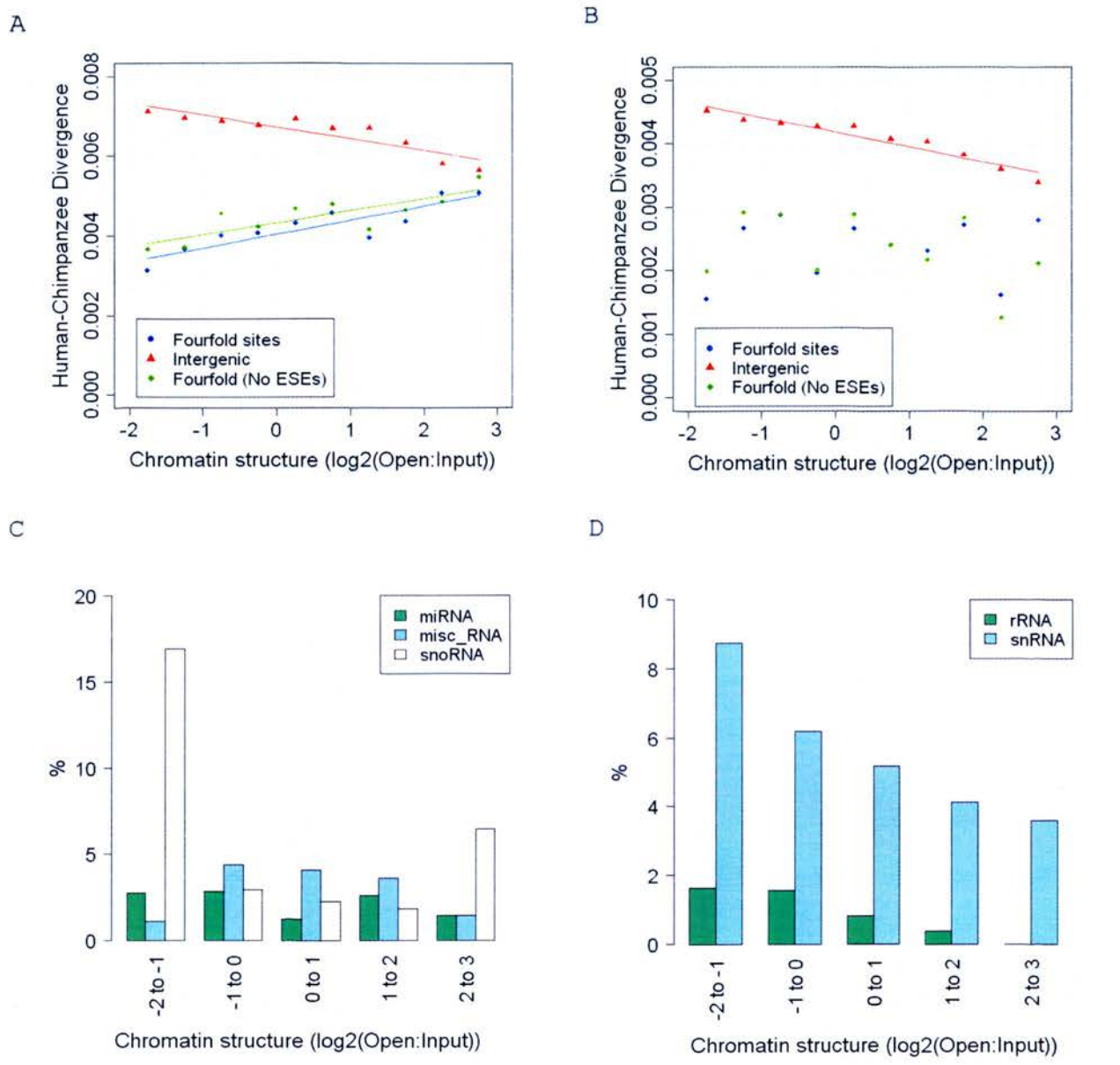


Figure 5

The effect of ESEs on fourfold degenerate site divergence and the ncRNA gene distributions observed across chromatin categories. (A+B) The observed rate of transitions and transversions respectively, at fourfold degenerate sites with and without ESE sites excluded (Fourfold degenerate site transversions r^2: = 0.02, p = 0.69; fourfold degenerate site transversions at non-ESE sites r^2: = 0.13, p = 0.30; intergenic transversions r^2: = 0.92, p = 1.2e-05. Fourfold degenerate site transitions r^2: = 0.78, p = 0.001; fourfold degenerate site transitions at non-ESE sites r^2: = 0.67, p = 0.004; intergenic transitions r^2: = 0.81, p = 4.0e-04). (C+D) The percentage of genes in each chromatin category that are of each Ensembl ncRNA class. Only the distributions of rRNAs and snRNAs show a significant negative correlation with chromatin structure (rRNA r^2: = 0.96, p = 0.004; snRNA r^2: = 0.92, p = 0.01)

mutation rate may be appropriate for a large proportion of genes. However, the use of dS as a surrogate measure of mutation rate for genes in closed chromatin will lead to the under-estimation of the true mutation rate in these regions and also the miscalculation of the levels of selection when used to measure dN/dS.

Exonic Splice Enhancers and RNA secondary structure

It has been proposed that synonymous sites may experience constraint because they play a role in controlling splicing or RNA stability [15]. For example, synonymous sites may be part of an exonic splice enhancer (ESE) motif or lead to a more stable base-paired RNA that is less susceptible to degradation. Although codon usage bias (resulting from unequal abundances of tRNAs and subsequent selection at synonymous sites in favour of codons corresponding to the most abundant tRNAs) has also been proposed as an explanation of synonymous site selection, the evidence for this in mammals is weak [19]. We therefore looked at the distribution of each predicted ESE motif across chromatin categories to see if their relative densities could explain the high levels of synonymous selection in closed chromatin. The density of a large proportion of ESE hexamers (44%) displayed a significant negative correlation with chromatin structure. However, given the base composition of ESE hexamers and coding regions across chromatin categories, we actually observed far fewer hexamers displaying a negative correlation than we would expect by chance (66%). This is because coding sequence base composition is itself correlated with chromatin structure and ESEs also show biases in their base composition. As shown in Figure 5A, excluding all sites from coding regions that overlap a predicted ESE hexamer leads to only a small increase in the rate of transitions observed at fourfold degenerate sites. Consequently, either there are many ESE motifs that are yet to be determined, or selection at synonymous sites is at most only partly the result of exonic splice enhancers.

We also compared the distribution of gene types across chromatin categories. If genes whose RNA structure is important were preferentially located in closed chromatin we may expect an over-representation of non-protein coding genes in closed regions. As shown in figures 5C+D, certain classes of non-protein coding genes are indeed over-represented in closed chromatin (rRNAs and snRNAs), while the distribution of other types of genes such as miRNAs and snoRNAs show no relationship with chromatin structure.

Further analysis is therefore required to determine why protein coding genes in closed regions of the genome display such comparatively high levels of selection at their synonymous sites. If it is indeed because of a requirement for a more stable secondary structure, then we may expect that the predicted stability of mRNAs from closed regions would be greater than those in open [20]. Future tests of this kind may help determine the reasons behind the enrichment of selection at synonymous sites in closed chromatin observed in this study.

Conclusion

We have shown that rates of mutation (intergenic, intronic and ancient repeat divergence as well as SNP density) and synonymous selection are correlated with chromatin structure. Regions of open chromatin display the lowest mutation rates and the least constraint at the synonymous sites of genes. Consequently previous observations of mutational hotspots in the human genome, high mutation rates around classes of genes involved in extracellular communication, the low dN/dS observed in housekeeping genes and the clustering of genes with similar divergence levels can all also be associated with chromatin structure. These correlations are observed despite the relatively low resolution of the chromatin dataset. The average length of the clones used in this analysis was 146 kb but the average human exon is approximately a thousand times smaller than this. There is consequently a disparity between the DNA regions whose rate of change we are measuring and the regions whose chromatin structure is known. The ability to measure chromatin structure at a higher resolution in the future may help increase the strength of these observed correlations.

We believe the lower background mutation rate observed in open regions of the genome in this study is likely to be a result of these regions being more accessible to repair mechanisms. Indeed it is known that sites of transcription-coupled repair are clustered in the gene dense (and therefore) open chromatin regions of the genome [21], that chromatin remodelling is a precursor to DNA repair, and that efficient DNA lesion detection is associated with relaxed chromatin structures [22-24]. However, contrary to mutation rate, we believe it unlikely that chromatin structure mediates selection on synonymous sites directly. Rather, it is more likely that genes that display a high level of selection at their synonymous sites are preferentially located in closed regions of the genome. It may be that these genes in general require especially tight transcriptional regulation, with a consequence being they are less accessible for DNA repair.

Chromatin structure is likely, however, to be only one of a number of factors that are associated with the variance in divergence rates observed across the human genome. This is supported by the fact that the levels of intergenic divergence of chromosome 19 are substantially higher than other autosomes, despite being gene dense and relatively open in structure. Most notably, both the chromatin dataset used in this analysis, as well as nucleosome forma-

tion potential [10], have previously been associated with GC content. Although this agreement between the lymphoblastoid chromatin dataset used in this analysis and other more general datasets is reassuring, GC content has previously been associated with rates of mutation and selection. However, although the mechanisms underlying the appearance of GC variability and isochores along the human genome remain controversial, it has been proposed that they may be a result of selection for the structural requirements of DNA. For example, an increase in GC content has been associated with an increase in bendability of DNA and a decrease in curvature, properties associated with more open chromatin [25]. Further analysis is consequently required to determine the complex interplay between the various factors involved in rates of mutation and selection across the human genome.

Methods

The abundance of open chromatin fibre structure in lymphoblastoid cells, at clones spaced approximately 1 Mb apart along the human genome, was determined as previously described [1]. Relative chromatin structure was represented in this analysis by $\log_2$(open chromatin:input chromatin) values (determined by cohybridising differentially labelled "open" and input chromatin fragments to a human genomic DNA microarray). A large $\log_2$(open:input) value in this analysis indicates a region enriched with open chromatin (see Gilbert et al. for further details). Clones with similar $\log_2$(open:input) values were binned for analysis (with bin sizes adapted to the amount of data available). The 2,787 human protein coding genes that mapped to each of these clones and their corresponding mouse orthologues were obtained from Ensembl (unique best reciprocal hits were taken where possible then reciprocal hits based on synteny). Coding sequence alignments of each of these orthologous pairs were derived via protein alignments (using the MUSCLE [26] and tranalign [27] programs). The codeml program of the PAML package [28] was used to calculate dN, dS and dN/dS using the F3 × 4 codon evolution model. Gene pairs with anomalously high dS values (> 1.270 i.e. twice the median dS of all human vs. mouse pairs) were excluded [29].

Gene expression breadth was determined through the analysis of the Gene Expression Atlas Affymetrix U133A dataset of Su et al. [30]. Intensity levels were averaged across arrays derived from the same tissue and all tumour derived arrays were excluded. A gene was defined as expressed if its mean signal level across all its corresponding probes exceeded that of the data set median [12]. To identify potential genes with CpG islands, the positions of predicted CpG clusters were obtained from the UCSC genome browser [31]. Of these islands, any that were less than 500 bp long, had a G+C content less than 55 or had a CpG to expected CpG ratio of less than 0.65 were excluded [32]. Those genes whose 5' end was within 2 kb of one of these islands were determined to be potential CpG island genes.

Human chimpanzee divergence was determined through the use of the chained and netted human-chimpanzee alignments available at the UCSC website (hg17-panTro1) [33]. Ensembl gene predictions were used to identify intronic, intergenic and protein coding regions. All exclusively intergenic and intronic regions found within clones were identified, and divergence measured in the corresponding sections of the human-chimpanzee alignment using PAML's baseml with the REV model [28]. Before calculating divergence all sequence from the same chromatin category was concatenated, in order to minimise the problems inherent in accurately measuring low divergence levels in regions of finite length. All bases that overlapped a CG dinucleotide in either species were removed from the alignments to conservatively calculate non-CpG rates of divergence [18].

Intergenic repeats were identified through UCSC's RepeatMasker annotation. Ancient repeats were defined as in Gibbs et al [29] and Taylor et al. [34] as repeats from the same RepeatMasker subfamily conserved between mouse and human in the same orientation. Simple repeats and regions of low complexity were excluded.

The SNP Consortium data were used to calculate SNP density across chromatin categories [35]. To ensure these densities were not biased as a result of the variety of protocols used to detect SNPs (some of which were chromosome specific), SNP densities across chromatin categories were also calculated using only SNPs randomly identified via the TSCM0019 protocol (a panel of 24 DNAs sequenced by the Sanger Centre, for more details see: [36]). The location of TSC SNPs was determined by mapping their ssIds to current rsIds via data available at dbSNP.

Predicted Exonic Splice Enhancer (ESE) hexamers were obtained from Fairbrother et al. [37]. The occurrence of each of these hexamers in the coding regions of each of the genes that mapped to a 1 Mb clone was determined. In order to identify the number of hexamers we would expect to detect by chance given the base composition of the genes and hexamers, we randomly shuffled the bases in each of the coding regions 100 times and recalculated the occurrence of each of the hexamers. The distribution of non-protein coding genes across chromatin categories was determined through Ensembl annotations.

Authors' contributions

JGDP undertook initial study design, software implementation, statistical analysis and interpretation, and drafted

the initial manuscript. NG and WAB determined the chromatin structure of the 1 Mb cloneset, participated in the study design and contributed to the final manuscript. HC, MGD and CAMS participated in the final study design, coordinated the study and contributed to the final manuscript.

Acknowledgements

JGDP is funded by an MRC Bioinformatics Research Studentship (G74/93). The work by MGD and HC is funded by grants from Cancer Research UK (C348/A3758), Medical Research Council (G0000657-53203) and Scottish Executive Chief Scientist Office (CZB/4/94). WAB, NG and CAMS are supported by the UK Medical Research Council. WAB is a James S.McDonnell Centennial fellow and is supported in part by FP6 through funding for the Epigenome Network of Excellence under contract LSHG-CT-2004-503433.

References

1. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA: **Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers.** *Cell* 2004, **118(5):**555-566.
2. Gazave E, Gautier P, Gilchrist S, Bickmore WA: **Does radial nuclear organisation influence DNA damage?** *Chromosome Res* 2005, **13(4):**377-388.
3. Arndt PF, Hwa T, Petrov DA: **Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects.** *J Mol Evol* 2005, **60(6):**748-763.
4. Chuang JH, Li H: **Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome.** *PLoS Biol* 2004, **2(2):**E29.
5. Williams EJ, Hurst LD: **proteins of linked genes evolve at similar rates.** *Nature* 2000, **407(6806):**900-903.
6. Lercher M, Chamary J, Hurst L: **Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile.** *Genome Res* 2004, **14:**1002-1013.
7. Sadoni N, Langer S, Fauth C, Bernardi G, Cremer T, Turner BM, Zink D: **Nuclear organization of mammalian genomes. Polar chromosome territories build up functionally distinct higher order compartments.** *J Cell Biol* 1999, **146(6):**1211-1226.
8. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent W, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey TS, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F, Haussler D: **Covariation in Frequencies of Substitution, Deletion, Transposition, and Recombination During Eutherian Evolution.** *Genome Res* 2003, **13:**13-26.
9. Lercher MJ, Hurst LD: **Human SNP variability and mutation rate are higher in regions of high recombination.** *Trends Genet* 2002, **18(7):**337-340.
10. Vinogradov AE: **Noncoding DNA, isochores and gene expression: nucleosome formation potential.** *Nucleic Acids Res* **33(2):**559-63.
11. Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A: **Patterns and rates of intron divergence between humans and chimpanzees.** *Genome Biol* 2007, **8(2):**R21.
12. Vinogradov AE: **"Genome design" model: evidence from conserved intronic sequence in human-mouse comparison.** *Genome Res* 2006, **16(3):**347-54.
13. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R: **Identifying biological themes within lists of genes with EASE.** *Genome Biology* 2003, **4(10):**R70.
14. Zhang L, Li W: **Mammalian housekeeping genes evolve more slowly than tissue-specific genes.** *Mol Biol Evol* 2004, **21(2):**236-239.
15. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev Genet* 2006, **7(2):**98-108.
16. Castresana J: **Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content.** *Nucleic Acids Res* 2002, **30(8):**1751-1756.
17. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci USA* 1993, **90(24):**11995-11999.
18. Chimpanzee Sequencing and Analysis Consortium: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437(7055):**69-87.
19. dos Reis M, Savva R, Wernisch L: **Solving the riddle of codon usage preferences: a test for translational selection.** *Nucleic Acids Res* 2004, **32(17):**5036-5044.
20. Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.** *Genome Biol* 2005, **6(9):**R75.
21. Surrallés J, Ramírez MJ, Marcos R, Natarajan AT, Mullenders LHF: **Clusters of transcription-coupled repair in the human genome.** *Proc Natl Acad Sci USA* 2002, **99(16):**10571-10574.
22. Gérard A, Polo SE, Roche D, Almouzni G: **Methods for studying chromatin assembly coupled to DNA repair.** *Methods Enzymol* 2006, **409:**358-374.
23. Loizou JI, Murr R, Finkbeiner MG, Sawan C, Wang ZQ, Herceg Z: **Epigenetic information in chromatin: the code of entry for DNA repair.** *Cell Cycle* 2006, **5(7):**696-701.
24. Rubbi CP, Milner J: **p53 is a chromatin accessibility factor for nucleotide excision repair of DNA damage.** *EMBO J* 2003, **22(4):**975-986.
25. Vinogradov AE: **Noncoding DNA, isochores and gene expression: nucleosome formation potential.** *Nucleic Acids Res* 2005, **33(2):**559-63.
26. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucl Acids Res* 2004, **32(5):**1792-1797.
27. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends in Genetics* 2000, **16(6):**276-277.
28. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.
29. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Ste-en D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera , Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bou-ard GG, Jong PJD, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Albà MM, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hübner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, López-Otín C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F, Consortium RGSP: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428(6982):**493-521.

30. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101(16):**6062-7.
31. Karolchik D, Baertsch R, Diekhans M, Furey T, Hinrichs A, Lu Y, Roskin K, Schwartz M, Sugnet C, Thomas D, Weber R, Haussler D, Kent W: **The UCSC Genome Browser Database.** *Nucl Acids Res* 2003, **31:**51-54.
32. Takai D, Jones PA: **Comprehensive analysis of CpG islands in human chromosomes 21 and 22.** *PNAS* 2002, **99(6):**3740-3745.
33. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci USA* 2003, **100(20):**11484-11489.
34. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM: **Heterotachy in mammalian promoter evolution.** *PLoS Genet* 2006, **2(4):**e30.
35. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Scha-ner S, Etten WJV, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, Group ISMW: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409(6822):**928-933.
36. **TSC TSCM0019 protocol** [http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewTable.cgi?method_id=581]
37. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive identification of exonic splicing enhancers in human genes.** *Science* 2002, **297(5583):**1007-1013.

Am. J. Hum. Genet. 77:112–119, 2005

Germline Susceptibility to Colorectal Cancer Due to Base-Excision Repair Gene Defects

Susan M. Farrington,[1,4,*] Albert Tenesa,[1,4,*] Rebecca Barnetson,[1,4] Alice Wiltshire,[1,4] James Prendergast,[1,4] Mary Porteous,[1,2] Harry Campbell,[1,3] and Malcolm G. Dunlop[1,2]

[1]Colon Cancer Genetics Group, School of Clinical and Molecular Medicine, [2]Clinical Genetics Department, and [3]Public Health Sciences, University of Edinburgh, and [4]Medical Research Council (MRC), Human Genetics Unit, Edinburgh

DNA repair is a key process in the maintenance of genome integrity. Here, we present a large, systematically collected population-based association study (2,239 cases; 1,845 controls) that explores the contribution to colorectal cancer incidence of inherited defects in base-excision repair (BER) genes. We show that biallelic *MUTYH* defects impart a 93-fold (95% CI 42–213) excess risk of colorectal cancer, which accounts for 0.8% of cases aged <55 years and 0.54% of the entire cohort. Penetrance for homozygous carriers was almost complete by age 60 years. Significantly more biallelic carriers had coexisting adenomatous polyps. However, notably, 36% of biallelic carriers had no polyps. Three patients with heterozygous *MUTYH* defects carried monoallelic mutations in other BER genes (*OGG1* and *MTH1*). Recessive inheritance accounted for the elevated risk for those aged <55 years. However, there was also a 1.68-fold (95% CI 1.07–2.95) excess risk for heterozygous carriers aged >55 years, with a population attributable risk in this age group of 0.93% (95% CI 0%–2.0%). These data provide the strongest evidence to date for a causative role of BER defects in colorectal cancer etiology and show, to our knowledge for the first time, that heterozygous *MUTYH* mutations predispose to colorectal cancer later in life. These findings have clinical relevance for BER gene testing for patients with colorectal cancer and for genetic counseling of their relatives.

Introduction

The role of base-excision repair (BER) in the maintenance of genome stability is primarily to counter oxidative DNA damage, which generates 8-oxoguanine products (8-oxoG). In humans, MYH (MIM 604933), OGG1 (MIM 601982), and MTH1 (MIM 600312) function in consort to identify and repair 8-oxoG incorporated into DNA, as well as to remove modified nucleoside. Recent studies have identified biallelic germline defects in *MUTYH*, in a proportion of families with multiple colorectal polyposis, that are not due to germline *APC* (MIM 175100 and 608456) mutations (Al-Tassan et al. 2002; Sieber et al. 2003). Although these studies provide indirect evidence, it is important to establish whether BER gene defects predispose to colorectal cancer (MIM 114500), to estimate the level of associated risk, and to determine the attributable contribution of such defects to overall disease incidence. Previous studies have provided some supporting evidence that biallelic mutations are associated with excess cancer risk (Croitoru et al. 2004; Fleischmann et al. 2004). Here, we present an analysis of the largest cohort study to date, thereby affording the opportunity to assess the effect of homozygous and heterozygous BER gene mutations on colorectal cancer risk. We assembled a large, systematically recruited prospective cohort of patients from across Scotland, shortly after diagnosis of colorectal cancer and irrespective of family history. We also systematically recruited healthy Scottish population control individuals through the central National Health Service (NHS). Using this population-based resource, we set out to determine, by a genetic association strategy, the role of BER genes in colorectal cancer susceptibility.

Received January 18, 2005; accepted for publication April 14, 2005; electronically published May 3, 2005.

Address for correspondence and reprints: Dr. Susan M, Farrington, Colon Cancer Genetics Group, MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom. E-mail: Susan.Farrington@hgu.mrc.ac.uk

* These two authors contributed equally to this work.

0002-9297/2005/7701-0011$15.00

Subjects and Methods

Assembly of the Cohort and Sample Collection

A populationwide accrual of prospective colorectal cancer cases has been in progress since 1999. Cases are ascertained through direct contact with every surgical and pathology department in Scotland. All cases had histologically confirmed adenocarcinoma of the colon or rectum. Blood DNA samples were obtained from patients after counseling and receipt of informed consent. Population-based and age- and sex-matched controls were

systematically identified, and blood DNA samples were obtained from them. A questionnaire about family information and lifestyle and medical history was completed for patients and controls. Dietary risk-exposure data were also collected by use of a validated food-frequency questionnaire. For cases, tumor stage, pathology, clinical presentation, as well as the presence of synchronous polyps were documented. These studies are subject to all necessary approvals from local ethics research committee (LREC), multicenter research ethics committee (MREC), and NHS research and development management in every participating hospital.

Analysis of MUTYH *Variants Y165C and G382D*

A two-stage approach was used to efficiently identify subjects carrying heterozygous or homozygous BER gene variants. In the first step, assays were designed using Primer Express v2.0 software (Applied Biosystems [AB]) for the two commonly reported variants, Y165C and G382D, in patients with multiple polyposis. Allelic discrimination for each variant employed allele-specific *Taq*Man MGB probes (AB), resolved on an ABI 7900 Analyzer, by use of SDS v2.1 software. Probe and primer details are available on request. Each Y165C and G382D variant identified by the *Taq*Man approach was confirmed by repeat DNA sequence analysis. In the second phase, all subjects heterozygous for Y165C and G382D underwent sequence analysis of the entire coding region of *MUTYH*, and the heterozygote cases were also screened for *OGG1* and *MTH1* gene variants. GenBank accession numbers for the genes are AF527839, NT_022517, and NT_007819, respectively. PCR products were treated with Exonuclease 1, shrimp alkaline phosphatase (Amersham Biosciences) and then were sequenced using ABI Big Dye terminator V3.0 chemistry, with precipitated products separated on an ABI 3700. Details of primer sequences are available on request. Using this approach, we identified all homozygotes or compound heterozygotes and any subjects in whom at least one allele was either Y165C or G382D. However, it should be noted that we could have missed some biallelic carriers who did not have these two common variants. Hence, if anything, our estimates may underrepresent the contribution of BER genes, but we consider this to be a marginal effect.

Assessment of Variants of BER Genes

Each variant identified by sequencing was confirmed by repeat sequence analysis. Allele frequency for each newly identified variant was then determined in at least 340 control chromosomes, to confirm that the variant was not simply a common polymorphism. Any common polymorphic variants were discarded, after which each identified variant was subjected to rigorous bioinformatic analysis. We used SIFT, which predicts deleterious coding variants on the basis of cross-species conservation; PolyPhen, which predicts the effect by use of conservation and any protein structure available in the public domain; T-Coffee, which aligns closely related Ensembl orthologues; protein domains predicted using Pfam; assessment of potential splicing effects by use of GENSCAN; ESEfinder; and Berkeley *Drosophila* Genome Project splice-site prediction.

Functional Analysis of MUTYH *nt 9639 a→g Variant*

A variant was identified that was predicted to affect splicing (*MUTYH*; nt 9639 a→g). In this case, RNA was extracted from blood leukocytes by use of Tri-reagent (Sigma) and was processed to cDNA (Boeringer Mannheim First Strand Synthesis kit). A transcript product of 475 bp was amplified using exonic primers in exons 10 and 14 (primer details available on request). The product was cloned using the Topo cloning kit (Invitrogen), and positive colonies were amplified and sequenced. Similarly, a genomic amplicon covering the nt 9639 a→g change and the G382D locus was cloned to demonstrate recessive inheritance of the two variants.

Test for Association

Association was tested using Fisher's exact test in a 3 × 2 table of genotype by colorectal cancer status, thereby making no prejudgment on potential mode of inheritance.

Estimate of Genotype Relative Risk

As an adaptation of the method of Hugot et al. (2001), the genotype relative risk (GRR) is defined as follows:

$$\mathrm{GRR}(AA) = \frac{\Pr(D|AA)}{\Pr(D|AA)} = 1 ,$$

$$\mathrm{GRR}(Aa) = \frac{\Pr(D|Aa)}{\Pr(D|AA)} = \frac{\Pr(Aa|D) \times \Pr(AA)}{\Pr(AA|D) \times \Pr(Aa)} ,$$

and

$$\mathrm{GRR}(aa) = \frac{\Pr(D|aa)}{\Pr(D|AA)} = \frac{\Pr(aa|D) \times \Pr(AA)}{\Pr(AA|D) \times \Pr(aa)} ,$$

where $\Pr(G|D)$ is the frequency of genotype G among cases and $\Pr(G)$ is the expected Hardy-Weinberg equilibrium proportion, as obtained from the allele frequency in the control population. The control population was assumed to be a representative sample of the general population. The 95% CIs for the GRRs were obtained

Table 1

Population GRR Associated with MYH Y165C

Population and Genotype	No. of Subjects with Genotype: Case	Control	GRR	95% CI
Entire cohort (P = .228):				
GG	0	0	.00	.00–.00
AG	17	8	1.76	.90–4.16
AA	2,202	1,824	1.00	1.00–1.00
U55 (P = 1.000):				
GG	0	0	.00	.00–.00
AG	7	4	1.08	.37–3.72
AA	867	535	1.00	1.00–1.00
O55 (P = .180):				
GG	0	0	.00	.00–.00
AG	10	4	2.42	.96–8.71
AA	1,335	1,289	1.00	1.00–1.00

by bootstrapping 1,000 independent samples. For each of the samples, the GRR was estimated, then the same 1,000 samples were ordered within each genotype, and the 50th and 950th estimates were taken as the lower and upper limits of the 95% CI.

Penetrance Estimates

Penetrance for colorectal cancer at a given age is defined as the probability that a randomly selected individual with genotype G will develop the disease by that age, with the assumption that that individual does not die of another cause. If A_x is the event "affected at age $x{:}x+1$, given being disease-free at x," then the probability that an individual is affected by age x, given the genotype G, is $\Pr(A_x|G)$. Similarly if $\overline{A}_x$ is the event "not affected at age $x{:}x+1$, given being disease-free at x," then the probability that an individual is not affected by age x, given genotype G, is $\Pr(\overline{A}_x|G)$. Population incidence rates, $\Pr(A_x)$, were obtained from Scottish Health Statistics, and the sex-average incidence rates were used.

Then

$$\frac{\Pr(A_x|G)}{\Pr(\overline{A}_x|G)} = \frac{\Pr(G|A_x) \times \Pr(A_x)}{\Pr(G|\overline{A}_x) \times \Pr(\overline{A}_x)} ,$$

$$\frac{\Pr(A_x|G)}{1-\Pr(A_x|G)} = \frac{\Pr(G|A_x) \times \Pr(A_x)}{\Pr(G|\overline{A}_x) \times [1-\Pr(A_x)]} ,$$

and

$$\Pr(A_x|G) = \frac{\frac{\Pr(G|A_x)\times\Pr(A_x)}{\Pr(G|\overline{A}_x)\times[1-\Pr(A_x)]}}{1+\frac{\Pr(G|A_x)\times\Pr(A_x)}{\Pr(G|\overline{A}_x)\times[1-\Pr(A_x)]}} .$$

There were not enough observations for each age, so $\Pr(G|A_x)$ was estimated using logistic regression (i.e., age was used as a predictor of genotype). Similarly, $\Pr(G|\overline{A}_x)$ was estimated from controls. The penetrance for genotype G is then defined as:

$$P_G = 1 - \prod_1^x [1 - \Pr(A_x|G)] .$$

Estimate of CIs

The 95% CI for the estimate of the penetrance was obtained by bootstrapping (i.e., by sampling with replacement) samples of cases and controls. A total of 500 independent samples were obtained of size equal to the total number of cases and controls. Penetrance was estimated for each sample, and mean penetrance over rep-

Table 2

Population GRR Associated with MYH G382D

Population and Genotype	No. of Subjects with Genotype: Case	Control	GRR	95% CI
Entire cohort (P = .010):				
AA	8	0	121.23	44.48–325.10
GA	35	20	1.46	.93–2.43
GG	2,181	1,808	1.00	1.00–1.00
U55 (P = .402):				
AA	4	0	146.69	26.71–998.71
GA	12	6	1.24	.54–3.42
GG	864	531	1.00	1.00–1.00
O55 (P = .049):				
AA	4	0	102.19	22.09–333.10
GA	23	14	1.60	.93–2.95
GG	1,317	1,277	1.00	1.00–1.00

Table 3

Population GRR Associated with *MUTYH* Gene

Population and Genotype[a]	No. of Subjects with Genotype		GRR	95% CI
	Case	Control		
Entire cohort (P = .001):				
MM	12	0	92.65	41.60–213.20
WM	45	28	1.35	.92–2.07
WW	2,160	1794	1.00	1.00–1.00
U55 (P = .976):				
MM	7	0	91.73	22.53–293.41
WM	14	10	.87	.43–1.72
WW	851	523	1.00	1.00–1.00
O55 (P = .014):				
MM	5	0	77.26	20.51–208.98
WM	31	18	1.68	1.07–2.95
WW	1,309	1,271	1.00	1.00–1.00

[a] W = wild-type allele; M = mutant allele.

licates and standard deviation (SD) were obtained. The 95% CI was calculated using the mean ± 1.96 SD.

Results

Analysis of MUTYH *Variants Y165C and G382D*

All variants detected by *Taq*Man analysis were confirmed by genomic sequencing. Eight (0.36%) of the patients were homozygous for G382D. There were no homozygotes for Y165C, but three Y165C/G382D compound heterozygotes were identified. There were no biallelic defects in any control samples. Association was first tested, using Fisher's exact test in a 3 × 2 table of genotype by colorectal cancer status, thereby making no prejudgment on potential mode of inheritance. Next, GRR was calculated using the method of Hugot et al. (2001), because the usual method, estimated by odds ratio, is not possible because there were no homozygote controls. The frequency of Y165C and G382D alleles in cases and controls is presented in tables 1 and 2, respectively.

The data presented in table 2 establish the fact that the G382D locus is associated with colorectal cancer in the complete cohort (P = .0104; GRR 121; 95% CI 44–325 for the G382D/G382D genotype). To explore any age-specific effect, the cohort aged <55 years at diagnosis (U55) was compared with those aged >55 years at diagnosis (O55) (table 2). Because of the lower number of U55 subjects, there was no statistically significant association with G382D for the U55 cohort, whereas association for the O55 cohort remained statistically significant (P = .0494). There was no association with the Y165C locus (table 1), possibly because of the rarity of the mutant variant in the Scottish population. We consider the lack of association in the early-onset cohort to be due primarily to the low allele frequency of these individual variants and the resultant lack of statistical power, even though this study involved large case and control cohorts.

Significance of Other Variants in BER Genes

In light of prior genetic and functional evidence of an additive effect of G382D and Y165C alleles (Al-Tassan et al. 2002; Sieber et al. 2003), we next determined the overall prevalence of biallelic defects, using a pragmatic approach. To find second BER gene defects, DNA from all carriers of monoallelic Y165C or G382D mutations (n = 74) was sequenced for each exon and intron/exon boundary of *MUTYH*. BER genes *OGG1* and *MTH1* were also sequenced in cases with heterozygous Y165C or G382D MYH alleles. All identified variants were excluded as polymorphisms by confirmation of wild-type sequence in at least 340 control chromosomes and likely functional relevance assessed by bioinformatic analysis, as described in the "Subjects and Methods" section.

In addition to the eight G382D/G382D homozygotes and the three Y165C/G382D compound heterozygotes described above, we identified a further patient with a heterozygous G382D mutation who had a second genomic *MUTYH* defect (nt 9639 a→g) that we consider to be pathogenic. This variant was confirmed to reside on the wild-type allele by genomic DNA cloning and sequencing. The variant was predicted to affect splicing. This was confirmed by cDNA analysis, which showed only mutant G382D transcript and no wild-type transcript. Thus, the patient was hemizygous for G382D at the RNA level. The *Taq*Man approach used for the G382D variant identified all nt 9639 a→g variants, so we have systematically screened all samples for this variant. Thus, the genotyping strategy would have identified the *MUTYH* defect nt 9639 a→g variant with equal sensitivity in cases and controls. Hence, we have included

this variant in further analyses of the effect of overall *MUTYH* genotype, for a total of 12 biallelic carriers (all cancer patients) and 73 subjects (45 patients and 28 controls) with monoallelic *MUTYH* alleles that we are confident are pathogenic (table 3).

We next considered other samples that might have a second BER gene defect. In all, there were an additional three samples with heterozygous G382D or Y165C alleles that had another BER gene defect that we consider likely to be pathogenic, but we cannot confirm or refute this. Hence, we did not include these other BER alleles in the analysis of the overall effect of putative BER gene defects (table 3). Two patients with Y165C or G382D mutations (one of each) had a second allelic variant in MTH1 (R31Q) that has been reported elsewhere in multiple polyposis (Sieber et al. 2003). Codon 31 of MTH1 is evolutionarily conserved, and bioinformatics interrogation suggests that the R31Q mutation affects protein function. Furthermore, we identified the variant in a total of 2 of 84 patient chromosomes, compared with 0 of 368 control chromosomes, which suggests that this is not a common polymorphic variant in the population (although we cannot directly compare R31Q prevalence, because these are two different groups). Taken together with previously published studies, we consider that the MTH1 R31Q variant is likely to be functionally important and consequently is likely to be a pathogenic mutation. One patient with an MTH1 R31Q variant also had a P391L variant in MYH that is at a highly conserved residue and is predicted to affect protein function. However, the significance of this variant is unclear, in light of the above discussion. A third patient with a monoallelic MYH G382D allele also had a variant in OGG1 (R197W). The variant is at a highly conserved codon, and bioinformatics analysis predicts a profound effect on protein function; again, the significance of this cannot currently be determined.

Analysis of the MUTYH *Gene and Colorectal Cancer Risk*

To assess biallelic inactivation of the entire gene, we calculated the combined risk for all significant *MUTYH* variants that had been analyzed in the complete cohort. This analysis included eight G382D/G382D homozygotes, three Y165C/G382D compound heterozygotes, and one nt 9639 a→g /G382D compound heterozygote, as discussed above. The data presented in table 3 convincingly establish the fact that the *MUTYH* gene is significantly associated with colorectal cancer ($P = .0012$). Biallelic inactivation imparts an overall 93-fold excess risk (GRR 93; 95% CI 42–213). As was the case for analysis of the G382D variant alone, the reduction in numbers for the U55 cohort resulted in loss of statistical power to detect association with the gene in that group, although the O55 group remains significant ($P = .0138$).

Separation of the cohorts by age reveals an age-specific risk effect and shows, for the first time, a significant monoallelic effect for late-onset disease when the empirical CIs produced by bootstrap analysis are used. There was a 1.68-fold excess risk (95% CI 1.07%–2.95) for heterozygous carriers aged >55 years and a population-attributable risk in this age group of 0.93% (95% CI 0%–2.0%). However, these results should be taken with some caution, because the significance was marginal, despite the fact that we have studied large case/control cohorts. Furthermore, analysis by use of standard CIs for the odd ratios in testing for association gave an almost significant result ($P = .085$; 95% CI 0.93–3) for a monoallelic effect. It was only with bootstrap analysis that this effect was significant at the 5% level. Nonetheless, these are novel observations that suggest that heterozygous *MUTYH* variants impart a modest increased risk later in life.

Estimate of Penetrance for MUTYH *Gene Defects*

We first estimated penetrance for G382D alleles at the *MUTYH* locus using the method of Satagopan et al. (2001), modified for application to a recessive trait. Figure 1 shows age-specific penetrance and 95% CIs for the G382D/G382D genotype alone. Of G382D/G382D carriers, ~55% developed colorectal cancer by age 40 years, and cancer had developed in all G382D/G382D carriers by age 65 years. Next, we estimated age-specific penetrance for all systematically determined biallelic defects identified in *MUTYH* (fig. 2). This analysis suggests that biallelic inactivation of the gene is highly penetrant, with all homozygous carriers developing cancer by age 60 years.

These findings establish the fact that there is a substantially elevated colorectal cancer risk early in life for people with biallelic MYH defects.

Synchronous Polyps in Biallelic MUTYH *Cancer Patients*

Data on polyp prevalence in cancer patients by genotype are presented in table 4. It is noteworthy that 4 (36%) of the 11 biallelic carriers with cancer for whom we had reliable polyp information had no concurrent metaplastic or adenomatous polyps. This emphasizes the fact that using polyp presence as a surrogate to enrich for *MUTYH* mutation carriers is quite insensitive. The data presented in table 4 show that there was a clear relationship between synchronous polyps and *MUTYH* genotype (all polyp types $P = .025$; adenomatous polyps $P = .007$; multiple adenomas $P < .0001$ [with use of Fisher's Exact Test]). In all, 64% of biallelic carriers had polyps; in all cases, these were multiple adenomas.

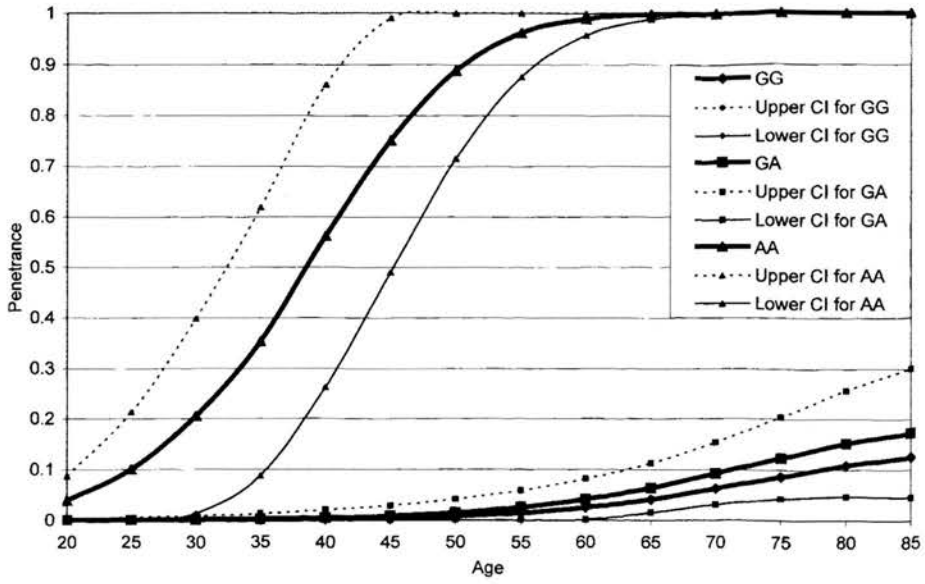


Figure 1 Penetrance curve of the MYH G382D locus. Age-related penetrance was estimated and plotted for all three genotypes, including 95% CIs.

The same analyses were performed, but with comparison of only the heterozygotes and wild-type genotypes, using 2 × 2 contingency tables; there were no significant results for presence of polyps with a heterozygote genotype ($P > .24$).

Discussion

Taken together, these data establish conclusively that defects in BER genes predispose to colorectal cancer. The findings provide robust evidence of causal involvement of biallelic BER gene defects in early-onset colorectal cancer, given the fact that there was complete penetrance by age 60 years. These findings substantially extend previous indirect evidence of involvement in colorectal cancer susceptibility through study of families with multiple polyposis (Al-Tassan et al. 2002; Sieber et al. 2003) and, combined with our findings of an excess of adenomatous polyps, suggest that polyps in such cases are premalignant. Our findings are also supported by previous suggestive evidence from smaller studies that lacked power to definitively establish whether there is an association between biallelic *MUTYH* variants and colorectal cancer (Enholm et al. 2003; Fleischmann et al. 2004; Wang et al. 2004).

There is only one previous study (Croitoru et al. 2004) that has shown an association between biallelic *MUTYH* variants and colorectal cancer. In that study, the authors proposed indirect evidence of a heterozygous effect because they observed nonrandom loss of heterozygosity of the wild-type allele in colorectal cancer. The findings presented here do, in fact, support their view and provide the first evidence of a modest colorectal cancer risk associated with heterozygous *MUTYH* mutations later in life. It is noteworthy that we found evidence of a significant excess risk associated with heterozygous *MUTYH* gene defects only when we considered the entire gene effect, and, importantly, this effect emerged only for late-onset disease. The effect was small and only just significant, and the results should be taken with some caution. Bootstrapping CIs are generally more robust than asymptotic CIs; however, it must be noted that asymptotic CIs for the odds ratio gave a slightly less significant result for the same data set. It is also possible that some of the excess risk we are detecting in heterozygotes is due to variants on the other allele that we have not detected. Data from mouse models indicates that, on an $Apc^{Min/+}$ background, monoallelic *Myh* inactivation does not increase tumor burden or the signature G-T transversions of the remaining *Apc* allele in the mouse tumors (Sieber et al. 2004). However, the study presented here is the largest to date, and previous studies have concentrated on early-onset disease or mixed cohorts (Fleischmann et al. 2004; Wang et al. 2004). Thus, only studies involving even larger numbers of later-onset cancer cases might be able to confirm our evidence of a monoallelic effect.

The study presented here emphasizes the requirement for very large population-based cohorts for robust assessment of the role of putative colorectal cancer sus-

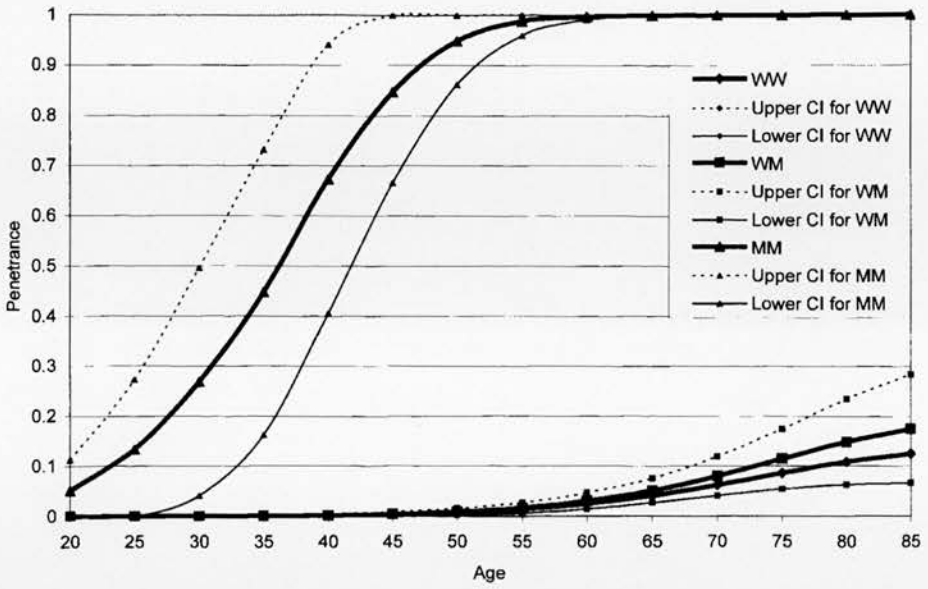


Figure 2 Penetrance curve of the *MUTYH* gene. All systematically verified and functionally deficient variants were included in the calculations, to estimate the penetrance of the entire gene. This therefore includes variants G382D, Y165C, and nt 9639 a→g. W = wild-type allele; M = mutant allele.

ceptibility alleles by genetic association strategies, even when relative risk is high. The relatively low allele frequency is likely to be a feature of many other cancer susceptibility alleles and raises some concerns for future power to detect associations where there is no prior biological hypothesis behind locus selection.

For clinical purposes, it is interesting to note that more than one-third of biallelic carriers did not have any synchronous polyps. This emphasizes the fact that use of polyps as an approach to enrich for *MUTYH* carriers with cancer is quite insensitive, in contrast with the findings of *APC*-negative multiple polyposis cases (Sampson et al. 2003; Sieber et al. 2003). Overall, there was an excess of synchronous polyps in biallelic carriers compared with others in this cohort, especially when adenomatous lesions were considered. Interestingly, monoallelic carriers did not appear to have an excess of polyps, suggesting the possibility that the excess cancer risk for heterozygotes later in life might be through a different mechanism.

Data presented here indicate that 1 of 50 patients who are diagnosed with colorectal cancer at age <40 years and 1 of 150 patients aged <55 years carry biallelic mutations in BER genes that are causally linked to cancer development. In defining the mutation carrier frequency for *MUTYH* and other BER genes, the present work informs future decisions about offering genetic testing to patients with early-onset colorectal cancer. Furthermore, the findings also have substantial clinical importance for the siblings of carriers, who have at least a 1/4 risk of colorectal cancer by age <60 years, by nature of the recessive genetic trait segregating in the family. Very large studies of late-onset disease-carrier status are required to replicate or refute the evidence

Table 4

***MUTYH* Genotype and the Presence of Synchronous Benign Polyps**

	No. (%) of Subjects with				
Genotype	No Polyps	All Polyp Types	Multiple Polyps	Adenomas	Multiple Adenomas
MM (n = 11)	4 (36)	7 (64)	7 (64)	7 (64)	7 (64)
WM (n = 44)	31 (70)	13 (30)	5 (11)	8 (18)	3 (7)
WW (n = 1,167)	859 (74)	308 (26)	100 (9)	224 (19)	54 (5)

Note.—Adenomas were either tubular or tubulovillous subtypes. Other polyps were metaplastic/hyperplastic. Three or more polyps were categorized as "multiple." Reliable concurrent polyp-prevalence data were not available for all patients with cancer, so the numbers for whom information was available are provided for each genotype.

presented here that monoallelic *MUTYH* defects contribute to colorectal cancer incidence.

Acknowledgments

We gratefully acknowledge the participation of all patients and control individuals, without whom this work would not have been possible. We thank all of the nursing and office staff employed by the Colorectal Cancer Genetic Susceptibility Study and the MRC Scottish Colorectal Cancer Study, for their tireless work in recruitment, which has been a major logistic endeavor. We also acknowledge the collaborative environment that we have enjoyed with NHS consultant surgeons and nursing teams in every Scottish hospital. We also thank the relevant departments in central Scottish NHS, including Cancer Registry, the Scottish Cancer Intelligence Unit of the Information and Statistics Division, and the Family Practitioner Committee, for invaluable help in recruiting population controls. We also thank Andrew Carothers for statistical advice. Current ethical approvals from MREC and LREC for all aspects of the study are held by M.P., H.C., and M.G.D. The work is funded by Chief Scientist Office grant CZB/4/94, Cancer Research UK grant C348/A3758, and MRC grant G0000657-53203. J.P. is funded by MRC Bioinformatics Research Studentship G74/93.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

Berkeley *Drosophila* Genome Project, http://www.fruitfly.org/seq_tools/splice.html
ESEfinder, http://rulai.cshl.edu/tools/ESE/
GenBank, http://www.ncbi.nlm.nih.gov/GenBank/ (for *MUTYH* [accession number AF527839], *OGG1* [accession number NT_022517], and *MTH1* [accession number NT_007819])
GENSCAN, http://genes.mit.edu/GENSCAN.html
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *MYH*, *OGG1*, *MTH1*, *APC*, and colorectal cancer)
Pfam, http://www.sanger.ac.uk/Software/Pfam/
PolyPhen, http://www.bork.embl-heidelberg.de/PolyPhen/
SIFT, http://blocks.fhcrc.org/%7Epauline/SIFT
T-Coffee, http://www.ch.embnet.org/software/TCoffee.html

References

Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR, Cheadle JP (2002) Inherited variants of *MYH* associated with somatic G:C→T:A mutations in colorectal tumors. Nat Genet 30:227–232

Croitoru ME, Cleary SP, Di Nicola N, Manno M, Selander T, Aronson M, Redston M, Cotterchio M, Knight J, Gryfe R, Gallinger S (2004) Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. J Natl Cancer Inst 96:1631–1634

Enholm S, Hienonen T, Suomalainen A, Lipton L, Tomlinson I, Karja V, Eskelinen M, Mecklin JP, Karhu A, Jarvinen HJ, Aaltonen LA (2003) Proportion and phenotype of *MYH*-associated colorectal neoplasia in a population-based series of Finnish colorectal cancer patients. Am J Pathol 163:827–832

Fleischmann C, Peto J, Cheadle J, Shah B, Sampson J, Houlston RS (2004) Comprehensive analysis of the contribution of germline *MYH* variation to early-onset colorectal cancer. Int J Cancer 109:554–558

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Sampson JR, Dolwani S, Jones S, Eccles D, Ellis A, Evans DG, Frayling I, Jordan S, Maher ER, Mak T, Maynard J, Pigatto F, Shaw J, Cheadle JP (2003) Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of *MYH*. Lancet 362:39–41

Satagopan JM, Offit K, Foulkes W, Robson ME, Wacholder S, Eng CM, Karp SE, Begg CB (2001) The lifetime risks of breast cancer in Ashkenazi Jewish carriers of *BRCA1* and *BRCA2* mutations. Cancer Epidemiol Biomarkers Prev 10: 467–473

Sieber OM, Howarth KM, Thirlwell C, Rowan A, Mandir N, Goodlad RA, Gilkar A, Spencer-Dene B, Stamp G, Johnson V, Silver A, Yang H, Miller JH, Ilyas M, Tomlinson IP (2004) *Myh* deficiency enhances intestinal tumorigenesis in multiple intestinal neoplasia ($Apc^{Min/+}$) mice. Cancer Res 64:8876–8881

Sieber OM, Lipton L, Crabtree M, Heinimann K, Fidalgo P, Phillips RK, Bisgaard ML, Orntoft TF, Aaltonen LA, Hodgson SV, Thomas HJ, Tomlinson IP (2003) Multiple colorectal adenomas, classic adenomatous polyposis, and germline mutations in *MYH*. N Engl J Med 348:791–799

Wang L, Baudhuin LM, Boardman LA, Steenblock KJ, Petersen GM, Halling KC, French AJ, Johnson RA, Burgart LJ, Rabe K, Lindor NM, Thibodeau SN (2004) *MYH* mutations in patients with attenuated and classic polyposis and with young-onset colorectal cancer without polyps. Gastroenterology 127:9–16